... / 500 Prototype 1

```
----create-DIR---
----before-curl---
total 925M
-rw-r--r-- 1 zeppelin hadoop  58M Feb  4 12:35 household_data_15min_singleindex.csv
-rw-r--r-- 1 zeppelin hadoop 851M Feb  3 17:57 household_data_1min_singleindex.csv
-rw-r--r-- 1 zeppelin hadoop  15M Feb  4 12:33 household_data_60min_singleindex.csv
----after-curl----
total 925M
-rw-r--r-- 1 zeppelin hadoop  58M Feb  4 12:35 household_data_15min_singleindex.csv
-rw-r--r-- 1 zeppelin hadoop 851M Feb  3 17:57 household_data_1min_singleindex.csv
-rw-r--r-- 1 zeppelin hadoop  15M Feb  4 12:33 household_data_60min_singleindex.csv
----hdfs----
Found 8 items
drwxr-xr-x   - zeppelin hdfs          0 2021-02-01 12:12 /user/zeppelin/.sparkStaging
drwxr-xr-x   - zeppelin hdfs          0 2021-02-02 09:18 /user/zeppelin/bank2
drwxr-xr-x   - zeppelin hdfs          0 2021-02-04 12:57 /user/zeppelin/conf
-rw-r--r--   1 zeppelin hdfs  892027611 2021-02-03 14:52 /user/zeppelin/household_data_1min_singleindex.cs
```

DataManipulation with PySpark

```
2.3.2.3.1.0.0-78
3.6
yarn
```

```
root
 |-- utc_timestamp: string (nullable = true)
 |-- cet_cest_timestamp: string (nullable = true)
 |-- DE_KN_industrial1_grid_import: string (nullable = true)
 |-- DE_KN_industrial1_pv_1: string (nullable = true)
 |-- DE_KN_industrial1_pv_2: string (nullable = true)
 |-- DE_KN_industrial2_grid_import: string (nullable = true)
 |-- DE_KN_industrial2_pv: string (nullable = true)
 |-- DE_KN_industrial2_storage_charge: string (nullable = true)
 |-- DE_KN_industrial2_storage_decharge: string (nullable = true)
 |-- DE_KN_industrial3_area_offices: string (nullable = true)
 |-- DE_KN_industrial3_area_room_1: string (nullable = true)
 |-- DE_KN_industrial3_area_room_2: string (nullable = true)
 |-- DE_KN_industrial3_area_room_3: string (nullable = true)
 |-- DE_KN_industrial3_area_room_4: string (nullable = true)
 |-- DE_KN_industrial3_compressor: string (nullable = true)
 |-- DE_KN_industrial3_cooling_aggregate: string (nullable = true)
```

“ETL”

I want to split the df into one df for industrial, one df for public, and one df for residential units.

Further, I want to transform from the wide format to a long format. I create an additional column indicating the number of the unit.

df_residential

unit	dishwasher	freezer	grid_import	heat_pump	pv	washing_mas