…ser / 200 Planning

Planning

- Data Manipulation: At first I loaded via curl into a Spark DF and tried to perform the data wrangling steps with PySpark in Zeppelin (see Prototype 1). However, it was quite laborious and challenging. Thus, I performed the steps on my local computer using Pandas in PyCharm.
- Data Load into GW, HDFS and Spark (see Prototype 2)
- Analyse the data and visualize insights.