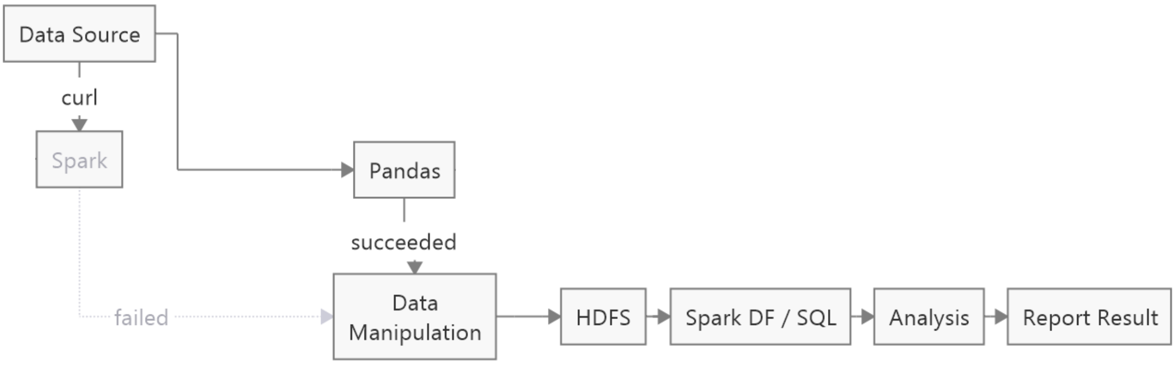
...ser / 400 Dataflow

Dataflow



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 and finally, I gave up.

Thus, I did the following procedure:
* Downloaded the data to my local computer.
* Manipulated the data using Pandas.
* Uploaded the prepared data to the gateway.
* Loaded the data into Spark.
* Analysed the data with Spark.
* Reported and visualized the results.