User-defined Machine Learning Functions

Markus Herrmann, Marc Fiedler
Global Data Science, GfK, Germany.

Abstract

In Data Science practices it is commonly assumed and accepted to abstract and slice big data architectures into functional layers, in particular a triad of governance-, data analysis- and persistence layer. However, moving input data to analysis, which is required when abstracting a data persistence layer from a data analysis layer, needs to be considered as highly expensive at large scale. Especially in Machine Learning (ML), the data analytics layer module requires intense data movements during preprocessing, data integration, preparation and analytics steps.

Therefore, we propose to consider an application of User-defined functions (UDFs) with ML capabilities directly at the data persistence layer, i.e. at the database. We observed that it might be overall most efficient in traditional on-premise (i.e. non-cloud) RDBMS environments to apply ML UDFs if only singular and self-contained ML tasks should be integrated.

Whereas the availability of ML functions in databases was predominantly owned by proprietary solutions in the past, there are now entirely new opportunities to integrate Python ML libraries with open source RDBMS. Whilst considering Python as one dominant language for ML applications in Data Science, the now achieved facilitation of Python ML UDFs consequently opens a broad range of opportunities to add Python ML capabilities to already existing persistence layers - without having to build an additional data analysis layer and related pipeline.

With this presentation we deliver preliminary results of our industry research about database centric ML applications, and we open source code for the application of (un)supervised learning models.

Keywords: *Machine Learning Engineering; RDBMS; UDF; MLUDF.*