IbmdbPy: Accelerating Python Analytics by In-Database Processing

Edouard Fouché and Michael Wurst

IBM Deutschland Research & Development GmbH

Abstract. The Python programming language is becoming widely used in data science and machine learning. Thus, Python ecosystem is very rich and provides intuitive tools for data analysis. However, most Python libraries require the data to be extracted from the database to working memory and ressources are limited by computational power and memory. Analyzing a large amount of data is often impractical or even impossible. IbmdbPy is an open-source python package, developed by IBM, which provides a Python interface for data manipulation and machine learning algorithms such as Kmeans or Linear Regression to make working with databases more efficient by seamlessly pushing operations written in Python into the underlying database for execution. This does not only lift the memory limit of Python, but also allows users to profit from performance-enhancing features of the underlying database management system. IbmdbPy is designed for IBM dashDB, a database system available on IBM BlueMix, the IBM cloud application development and analytics platform. Via remote connection, user operations can benefit from dashDB specific features, such as columnar technology and parallel processing, without having to interact with the database explicitly. Some in-database functions additionally use lazy loading to load only parts of the data that are actually required to further increase efficiency. Keeping the data in the database also avoids security issues that are associated with extracting data and ensures that the data that is being analyzed is as current as possible. IbmdbPy can be used by Python developers with very little additional knowledge, since it imitates the well-known interface of Pandas library for data manipulation and Scikit-learn library for machine learning algorithms. The project is still at an early stage, but several experiments have already been conducted to measure the advantage of using IbmdbPy over the corresponding in-memory implementation. The results show that it provides a great runtime advantage for operations on medium to large dataset, i.e. on tables that have 1 million rows or more. The project aims to extend the BlueMix ecosystem, by providing a Python interface for dashDB, bridging the gap between the analytics platform and end-user environment, so that developers can benefit both from the expressivity of Python and from the speed-up provided by SQL execution in dashDB, which can be run on a cluster.

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org