Hannes Mühleisen, CWI DA: *MonetDBLite – Bringing Column Stores to the Masses*

Data Science takes place in specialized scripting environments such as Python or R. These environments were never designed to handle huge datasets, but are now routinely used for precisely that task. Data management is an issue, reading large dataset from flat files or through a socket connection involves huge overheads and thus increases the time-to-analysis to impracticable amounts.

Analytical relational data management systems are well-suited to store a complex collection of tables and also provide transactional guarantees, but the installation and maintenance of a database server is a tall order for typical statisticians or domain researchers, for example geneticists.

We have recently published MonetDBLite [1], an in-process version of MonetDB, a free and Open Source relational database focused on analytical workloads. MonetDBLite requires no administration and installs through the normal package system of the host language. So far, we support R and Python as host languages. In addition, the MonetDB's columnar storage model translates well to R's vectors or Python's NumPy arrays. Therefore, data transfer to and from MonetDBLite tables are highly efficient. For example, transferring one Billion values from MonetDBLite into an R data.frame only requires 94 seconds on a standard laptop.

In our talk, we will present the concepts and innovations required to embed a highly complex data management system into a highly complex statistical scripting environment. We will also demonstrate the capabilities of MonetDBLite using real-world use cases from survey analysis.

1. H. Mühleisen: *MonetDBLite for R*, https://www.monetdb.org/blog/monetdblite-r