Supercharged Hybrids – Zero-Copy Database Integration for R

Hannes Mühleisen[1⋆] **and Jonathan Lajus**[2]

1. Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
2. ENS Cachan, Paris, France
⋆Contact author: hannes@cwi.nl

Keywords: Database Integration

Statistical analysts have long been struggling with ever-growing data volumes. While specialized data management systems such as relational databases would be able to handle the data volume, statistical analysis tools are far more capable to express and execute complex data analyses. A tight integration of these two classes of systems has the potential to overcome the data management issue while at the same time keeping analysis convenient. However, one must keep a careful eye on implementation overheads such as serialization, process switching and format translation.

There has been considerable interest in these so-called *hybrid* systems. For example, a decade after the initial work by Duncan Lang [4], both Oracle and SAP now sell in-database *R* environments [2, 1]. However, these integrations still pay considerable dues to overhead when transforming data back and forth between the database-native representation and R data types. However, modern analytical database systems rely on a columnar data representation as well [3], hence a very cheap or even "free" data transformation might be at hand.

In this talk, we present our work on the in-process integration of data management and analytical tools. In particular, we investigate under which conditions a *zero-copy integration* is feasible due to the omnipresence of C-style arrays containing native types. In this context, zero-copy refers to sharing data between a database and a statistical environment within the same process without a single modification. We discuss the general concept, its consequences and present a prototype of this integration based on the columnar relational database MonetDB and the *R* environment for statistical computing. We evaluate the performance of this prototype in a series of micro-benchmarks of common data management tasks. Our findings may be used to improve the *R* environment towards easier embedding.

References

[1] Große, P., W. Lehner, T. Weichert, F. Färber, and W.-S. Li (2011). Bridging two worlds with RICE integrating R into the SAP in-memory computing engine. *PVLDB 4*(12), 1307–1317.

[2] Hornick, M. and T. Plunkett (2013). *Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop*. McGraw-Hill Osborne Media.

[3] Idreos, S., F. Groffen, N. Nes, S. Manegold, K. S. Mullender, and M. L. Kersten (2012). MonetDB: Two decades of research in column-oriented database architectures. *IEEE Data Engineering Bulletin 35*(1), 40–45.

[4] Lang, D. T. (2001, 04). Scenarios for using R within a relational database management system server. `http://www.omegahat.org/RSPostgres/Scenarios.pdf`. checked 2013-10.

[5] Mühleisen, H. and T. Lumley (2013). Best of both worlds: relational databases and statistics. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, New York, NY, USA, pp. 32:1–32:4. ACM.