

# Managing data.frames with package 'ff'

Jens Oehlschlägel<sup>1,\*</sup>, Daniel Adler<sup>2</sup>

1. Truecluster.com, Munich
  2. Institute for Statistics and Econometrics, University of Göttingen
- \* Contact author: Jens\_Oehlschlaegel@truecluster.com

**Keywords:** Large data, databases, column stores

We explain the new capability of package 'ff 1.1' to store large dataframes on disk in class 'ffdf'. 'ffdf' objects have a virtual and a physical component. The virtual component defines a behavior like a standard dataframe, while the physical component can be organized to optimize the 'ffdf' object for different purposes: minimal creation time, quickest column access or quickest row access. Furthermore 'ffdf' can be defined without rownames, with in-RAM rownames or with on-disk rownames using a new 'ff' class 'fffc' for fixed width characters. On a standard notebook we give an online demo of processing an 80 mio row dataframe – size of a German census :-)

## References

- Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J. Zucchini, W. (2008-2009) R package ff 2.1.0 "Memory-efficient storage of large atomic vectors and arrays on disk and fast access functions for R"  
<http://cran.at.r-project.org/web/packages/ff/index.html>
- Adler, Oehlschlägel, Nenadic, Zucchini (2008) Large atomic data in R package 'ff'. Presentation at UseR!2008, statistics department, University of Dortmund
- Oehlschlägel, Adler, Nenadic, Zucchini (2008) A first glimpse into 'R.ff'. Presentation at UseR!2008, statistics department, University of Dortmund