Massively parallel analytics for large datasets in R with nza package

Przemyslaw Biecek[2,1*], Pawel Chudzian[1*], Cezary Dendek[1], Justin Lindsey[1]

1. Nzlabs, Netezza
2. Warsaw University

* Both authors contributed equally to this work.

Corresponding author: przemyslaw.biecek@gmail.com

Keywords: large datasets, parallel processing, high performance processing, Netezza Performance Server, data aggregates

One of the bottlenecks towards processing large datasets in *R* is the need of storing all data in memory. Therefore, users are limited to datasets that fit in 2 or 4 GB memory limit. To avoid this, the natural approach is to split statistical algorithms in two steps. In the first step the data processing is performed outside *R*, e.g. in database or on flat text file resulting in precomputed data aggregates. In the second step these aggregates are imported into *R* where the rest of the analysis is performed. Such data aggregates are called sufficient statistics, because they contain all information necessary to compute parameter estimates, test statistics, confidence intervals and model summaries while are much smaller than the original dataset.

Such an approach is implemented in the **nza** package. In the first step the **nzr** package is used to connect with Netezza Performance Server (NPS) and stored procedures are used to compute data aggregates in a parallel fashion. Having data stored in a parallel/multi-nodes database has two main advantages: there is no limit on the size of accessible datasets and data aggregates are computed in a parallel fashion which significantly reduces computation time. In some cases achieved reduction is linear with the number of processors in the database server.

Following algorithms are implemented in the **nza** package using sufficient statistic approach: correspondence analysis, canonical analysis, principal component analysis, linear models and mixed models, ridge regression, principle component regression and others.

In our presentation we will show the performance study for algorithms implemented in the **nza** package.