Parallel data processing with R on combined IBM Netezza 1000 and HPC cluster

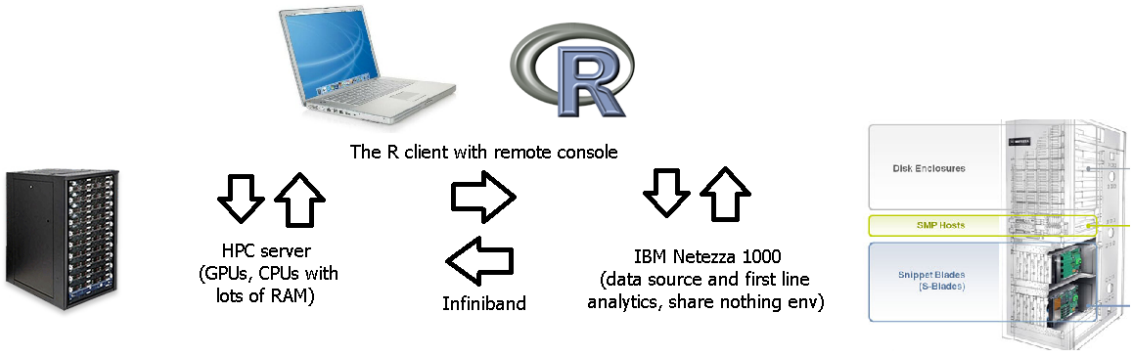
Przemyslaw Biecek*, Grzegorz Maj, Ala Strachocka, Pawel Chudzian, Maciej Michalewicz, David Flaxman

IBM Netezza R&D Labs, *Contact author: Przemyslaw.Biecek@pl.ibm.com

Keywords: IBM Netezza 1000, Netezza Performance Server, High Performance Cluster, parallel data processing

IBM Netezza 1000 is a warehouse that allows for using R functions in parallel in database very close to data. However in some cases user might want to offload the database and send computations to HPC server rich in RAM. In this talk we are going to present the prototype solution that benefits from parallel database IBM Netezza 1000 and the HPC server. Constructed framework facilitates implementing and executing parallel processing jobs. This is a next step to move analytics closer to big data and extend capabilities of in-database INZA procedures.

The main advantages of this approach are: reducing IBM Netezza 1000s overload by running computationally heavy tasks outside IBM Netezza 1000; more operational memory and faster processors on the cluster; dedicated hardware (e.g. GPU cards), since the cluster can be composed of diverse nodes; ultra-fast parallel data transfer via a dedicated connection.



- **Use case: Advanced analytics for Value at Risk in Operational Risk.** Data describing operational losses is stored in IBM Netezza 1000. Each row describes a loss, with corresponding risk category and business line. Modeling is done in parallel for each cell in the risk matrix. Data is distributed in IBM Netezza 1000 over risk categories and business lines and processed in a group-by-group fashion on the cluster. Each cluster node models losses distribution of a separate cell from the risk matrix.
- **Use case: Advanced Credit Scoring.** Both training and apply steps of the scoring process may be performed on the proposed architecture. Data describing credit or loans applications is stored in IBM Netezza 1000. In the first step different classifiers are trained on the whole training set. Model fitting and parameters tuning may be performed in parallel on different nodes. In the second step selected classifier or set of classifiers is applied in the row-by-row mode for all rows from the test/target dataset.
- **Use case: Bootstrap enhanced hierarchical genes clustering.** Data describing gene expressions is stored in the IBM Netezza 1000. The hierarchical clustering needs to be performed for bootstrap samples of the whole dataset. Thus the whole data set is distributed over computational nodes and the bootstrap operations are performed in parallel. The chunks of the dataset are sent to all computational nodes that broadcast their shares to all other nodes. Eventually, every node gathers the whole dataset. Bootstrap samples are generated and hierarchical clustering is executed on each node. Due to the high number of genes, each computational unit needs to be equipped with dozens of gigabytes of RAM. In the R statistical package different algorithms for hierarchical clustering are available, like fastcluster, which is order of magnitude times faster than standard implementation. Also, libraries that perform hierarchical clustering on GPU cards like gputools are available.