Analytics at Scale with *R*

Brian Hess[1], Michele Chambers[2]

1. Brian is Principal Mathematician and Director of Advanced Analytics at Netezza
2. Michele is Director of Advanced Analytics Product Management at Netezza

Keywords: In-database analytics, data mining, scoring, high performance computing, parallel algorithms

As *R* grows in commercial acceptance, companies are seeking ways to migrate their ad-hoc analysis into production deployment. Depending on the organization and the problem being addressed, there may be a desire to learn the model on data beyond a sample as well as applying the model to large scale data. Scaling up analytics takes the following basic forms:

- Data Intensity
    - Depth of data (ie: number of transactions, deeper level of transactions or more history)
    - Width of data (ie: number of factors, dimensions, features)
- Computational Intensity
    - Computational complexity (ie: k-means, PCA, heroic computations, matrix, linear algebra)
    - Model complexity (ie: number of experiments, simulations, more initial conditions)
- Parallel Intensity
    - Combination of computational intensity on data intensive problems

In order to address these emerging requirements, commercial companies supporting *R*, need to provide building blocks to make migration of *R* models and algorithms easier including but not limited to:

- Programming constructs (ie: iterators, foreach, etc.)
- Storage constructs (ie: striping data, etc.)

In this session, we want to engage with the *R* community to think about how best to separate the data layer from the processing layer in order to facilitate the commercial viability of *R* for large scale analytics.

References

http://www.netezza.com/data-warehouse-appliance-products/twinfin-i.aspx