Contribution ID: **0** Type: **not specified**

Introduction to R and in-database analytics using ORE

Tuesday 5 March 2013 10:10 (45 minutes)

R, http://cran.r-project.org is an open source project, which provides a programming language and environment for data manipulation, calculation, analysis and graphical display. Nowadays, due to many reasons such as, a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques, R is becoming a kind of the facto standard among statisticians and data miners for developing statistic and analytic software. R can be considered as an integrated framework that includes:

• Effective data handling and storage.
• Operators for high performance calculations on arrays and matrices.
• A large, coherent and integrated collection of intermediate tools for data analysis.
• Graphical facilities for data analysis and display either on-screen or on hardcopy.
• A well-developed, simple and effective programming language, which includes conditionals, loops, user-defined recursive functions and input and output facilities.

One of the most remarkable features in R is its highly extensible nature. R can be easily extended via packages. A huge number of packages ranging from simple statistic models to complex artificial intelligence techniques are available on the R repositories (CRAN).

In addition, recently Oracle has introduced the Oracle R Enterprise, ORE, as a component of the Oracle Advanced Analytics option. ORE is designed for problems involving large amounts of data and integrates R with Oracle databases. As much data you have as closer you want to perform your analysis form the database. Also ORE improves the original R capabilities in terms of parallelism and scalability, which in opinion of many R users represents the main issue for applying R based approaches on production services. ORE not only improve the aspects describe above but also introduced a real integration within Oracle databases. This integration is translated into the possibility to perform in-database analytics or, in other words, allows transparent access to the data stored in the database and therefore the possibility to execute embedded analysis or statistical computation. In an environment such as CERN where many of the fundamental components are database driven is an essential feature to improve the performance of the data analytics, which, at the same time, is also one of the central requirements for most of the analytics needs. The session will introduce some of the basic concepts of R, objects, looping functions, data visualization and graphical representation and finally we will review some more advance functionalities such as in-database analytics using ORE. Also some best practices and tools such as Rstudio or statET will be also introduced.

Presenter: MARTIN MARQUEZ, Manuel (CERN)