Revolution R Enterprise 7.0 README

Revolution R Enterprise 7.0 for 32-bit and 64-bit Windows and 64-bit Linux (Red Hat Enterprise Linux 5.x and 6.x and SUSE Linux Enterprise Server 11.x) features updated releases of the bundled RevoScaleR 'big data' package and the optional RevoDeployR package. A new package for interactively viewing decision trees, RevoTreeView, is also now bundled with Revolution R Enterprise

Installation instructions are provided in your welcome e-mail.

What's New in Revolution R Enterprise 7.0

Open Source R

R 3.0.2

- Revolution R Enterprise 7.0 provides the latest stable version of open source R: R 3.0.2

Documentation

Two new guides are provided to give an overview of Revolution R Enterprise 7.0:

- *What's New in Revolution R Enterprise 7.0*
- *Revolution R Enterprise 7.0 Getting Started Guide*

RevoScaleR

Documentation

Three new guides are provided to give information and examples on doing computations on specific distributed computing platforms:

- *RevoScaleR Hadoop Getting Started Guide*
- *RevoScaleR LSF Cluster Getting Started Guide*
- *RevoScaleR HPC Server Getting Started Guide*

Hadoop Compute Context

- With the RxHadoopMR compute context, you can distribute RevoScaleR computations across a Hadoop cluster (Cloudera CDH4 orHortonworks HDP 1.3 on RHEL5 or 6). Both High-Performance Analytics functions (such as rxSummary, rxLinMod, and rxLogit) and High-Performance Computing with rxExec are supported. Your data must be in HDFS in the form of .csv files or an .xdf file, and can be imported in Hadoop from a folder of .csv files into a composite .xdf. (See *the RevoScaleR Hadoop Getting Started Guide* for more information.)
- The RxHadoopMR compute context uses MapReduce behind the scenes, providing all of the advantages of MapReduce including scalability and fault tolerance. Analyses of large data sets will typically scale linearly with the number of nodes and linearly with the number of rows of data.
- Using MapReduce, there is an overhead of roughly 20-30 seconds for a single pass through the data. There is also an overhead of roughly 20-30 seconds for getting results from the cluster to the client laptop or workstation. So, typically it is more efficient to perform smaller analyses using a local compute context.

- Complete instructions for configuring your Hadoop cluster to work with Revolution R Enterprise are included in the *Revolution R Enterprise 7 Installation Guide for Linux System* (instman.pdf).

Decision Forests for 'Big Data'

- The new rxDecisionForest function allows estimation of an ensemble of decision trees.

Stepwise Regression for 'Big Data' Linear Models Using Significance Level Variable Selection

- The rxStepControl function has a new argument 'stepCriterion' which defaults to AIC but allows specification of 'stepCriterion="SigLevel"' to allow SAS-like stepwise regression where variables to be added or dropped are compared to the current model using F or Chi-squared tests of significance. Significance levels for entry to or removal from the model can be specified with the new arguments 'maxSigLevelToAdd' and 'minSigLevelToDrop'.

Stepwise Regression for 'Big Data' Logistic Regression and Generalized Linear Models

- The rxLogit and rxGlm functions have a new argument 'variableSelection' which allows you to specify variable selection criteria for stepwise model fitting. As with stepwise regression for linear models, the rxStepControl function can be used to provide the variable selection criteria, and the same variety of controls are available, including a choice between AIC and significance level selection criteria, ability to specify model scope, and choice of method among forward selection, backward elimination, or bi-directional search ('stepwise'). When analyzing big data sets, setting the 'refitEachStep' argument to FALSE will typically speed up the analysis considerably.

Duplicate Removal in rxSort

- The new argument 'removeDupKeys' allows you to remove duplicates while sorting. The first record with a unique combination of the 'sortByVars' is retained, subsequent records with the same combination are removed. If desired, a new variable can be created containing the number of records that match each combination. This can be used as a frequency weight in subsequent analyses.

Output to Text Data Files for rxDataStep and rxPredict

- It is now possible to specify a delimited text RxTextData data source as the outFile for rxDataStep and rxPredict when using a local compute context.

AIC Computations Added to rxGlm

- The rxGlm function has a new flag, computeAIC, that, if TRUE, causes the Akaike Information Criterion (AIC) to be computed and returned as part of the model.

Support for PMML Output

- New functions as.lm, as.glm, as.rpart, and as.kmeans convert appropriate RevoScaleR model objects to objects of class lm, glm, rpart, and kmeans, respectively. The underlying structure of the output object will be a subset of that produced by an equivalent call to the corresponding function. In many cases, this method can be used to coerce an object for use with the pmml package (available from CRAN or the Revolution Analytics versioned source repository).

Improved Data Path Handling

- The dataPath argument for rxOptions and data sources is now fully respected. The current working directory is not included in the data search path unless explicitly included with ".".

- The rxOptions function now contains an outDataPath argument to specify the default directory location for writing .xdf files from RevoScaleR functions.
- The local compute contexts RxLocalSeq, RxLocalParallel, and RxForeachDoPar contain two new slots, a dataPath slot for specifying the directory from which to read files used in RevoScaleR functions and a default outDataPath specifying the directory to which to write files from RevoScaleR functions.
- The distributed compute contexts RxHpcServer, RxAzureBurst, and RxLsfCluster, gain an outDataPath slot to specify the directory to which to write files from RevoScaleR functions. The dataPath and outDataPath for RxHadoopMR are not yet supported.

New Default for .xdf File Compression

- The default value for the xdfCompressionLevel, set in rxOptions, has been changed to 1 from 0. RevoScaleR functions that write .xdf files use this option to determine the default compression level. The previous default specified no compression; now compression is the default.
- The list returned by rxGetInfo now includes a compressionType component for .xdf files.

Data Sources Returned More Consistently

- The rxDataFrameToXdf function now returns an RxXdfData data source object representing the output .xdf file.
- The rxXdfToText function now returns an RxTextData data source object representing the output text file.

New Behavior When Importing Numeric Data as Logical

- When numeric data are imported as logical variables, any valid non-zero values are set to TRUE, to be more consistent with the standard R as.logical function.

RxAzureBurst Now Deprecated

RevoMPM

RevoMPM is the Revolution Multi-Node Package Manager. It can be used to manage Hadoop cluster installations. Instructions are provided in the *Revolution R Enterprise 7 Installation Guide for Linux Systems* (instman.pdf).

RevoTreeView

The new RevoTreeView package provides browser-based visualizations of decision trees fit with RevoScaleR's rxDTree function or the rpart package's rpart function. The createTreeView function creates the visualization, which can then be displayed in your default browser with the standard generic plot function; the zipTreeView function allows the visualization to be packaged for sharing with other users.

ParallelR

RevoScaleR is now the preferred package for parallel computing in Revolution R Enterprise. The foreach and iterators packages, together with the base R package parallel, provide additional parallel computing

facilities. Earlier approaches to parallel computing in R have now been removed from Revolution R Enterprise:

NetWorkSpaces Removed From Revolution R Enterprise

- NetWorkSpaces-based packages for parallel computing have been removed from Revolution R Enterprise 7.0. This includes the following packages:
    - nws
    - nwsserver
    - bootNWS
    - randomShrubberyNWS
    - sprngNWS
    - doNWS

  These packages remain available for download from http://sourceforge.net/projects/parallelr/

SNOW and multicore Packages Removed From Revolution R Enterprise

- The snow and multicore packages have been subsumed into the R base package parallel; the multicore package is now obsolete, per the R core team. The doMC package, which previously used multicore, has been updated to use the parallel package. The multicore and snow packages have been removed from Revolution R Enterprise 7.0; both packages are still available from CRAN.
- The doSNOW and doMC packages have been removed from Revolution R Enterprise; use doParallel instead. (Both doSNOW and doMC remain available from CRAN.)

Known Issues

- ***Known Issues in Revolution R Enterprise 7.0***