

Revolution R Enterprise

Portland R User Group

November 13, 2013

David Smith @revodavid

Michael Helbraun



**BIG
DATA**



DATA SCIENCE

OPEN SOURCE R



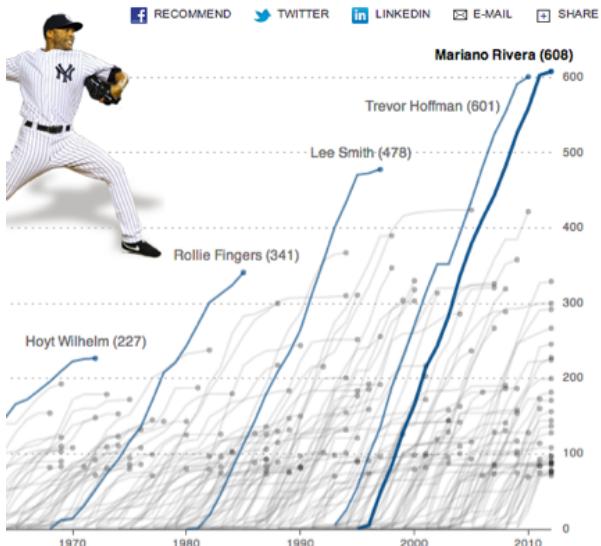


Innovate with R

- Most widely used data analysis software
 - Used by 2M+ data scientists, statisticians and analysts
- Most powerful statistical programming language
 - Flexible, extensible and comprehensive for productivity
- Create beautiful and unique data visualizations
 - As seen in New York Times, Twitter and Flowing Data
- Thriving open-source community
 - Leading edge of analytics research
- Fills the talent gap
 - New graduates prefer R

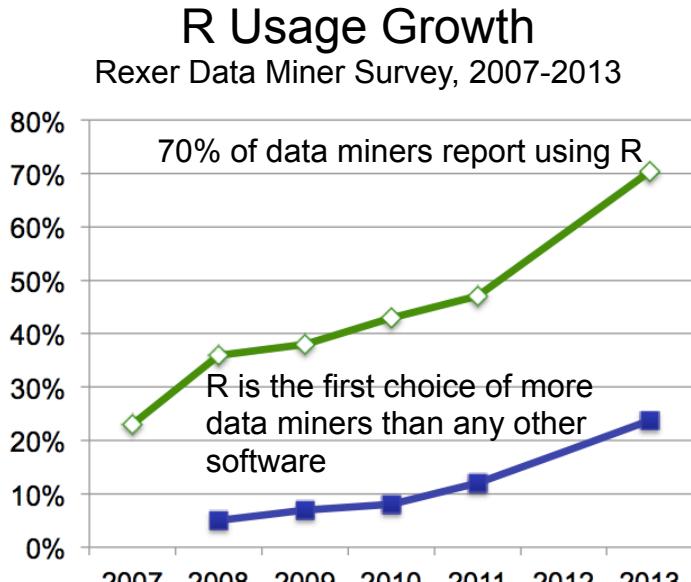
R is Hot
bit.ly/r-is-hot

WHITE PAPER





R is exploding in popularity & functionality



Source: www.rexeranalytics.com

"I've been astonished by the rate at which R has been adopted. Four years ago, everyone in my economics department [at the University of Chicago] was using Stata; now, as far as I can tell, R is the standard tool, and students learn it first."

Deputy Editor for New Products at Forbes

"A key benefit of R is that it provides near-instant availability of new and experimental methods created by its user base — without waiting for the development/release cycle of commercial software. SAS recognizes the value of R to our customer base..."

Product Marketing Manager SAS Institute, Inc



Revolution R Enterprise

Power R for the
Enterprise

Empower Platform
Independence



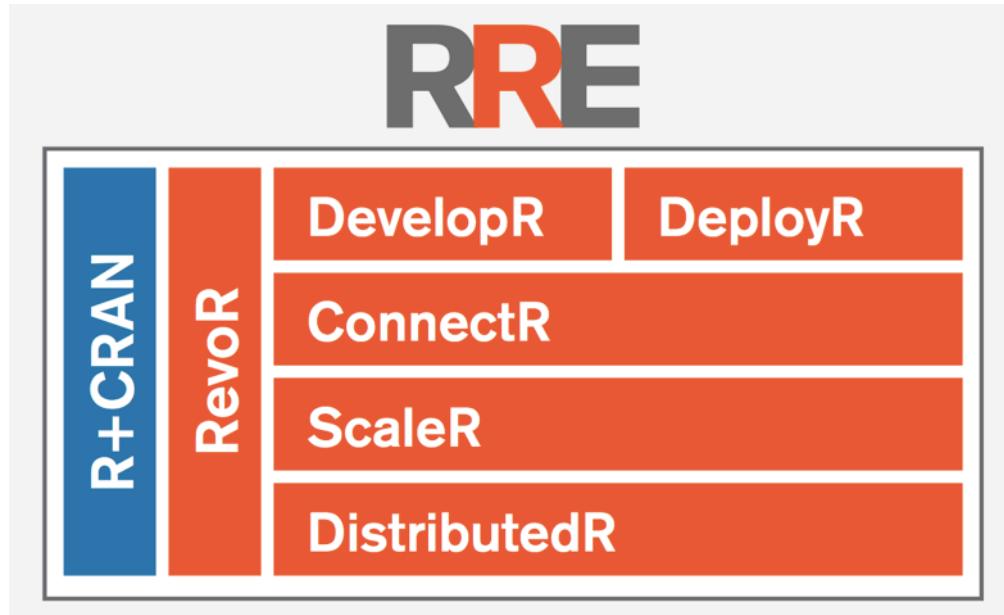
Supercharge R for
Massive Data

Take Big Cost Out
of Big Data





RRE is the Big Data Big Analytics Platform

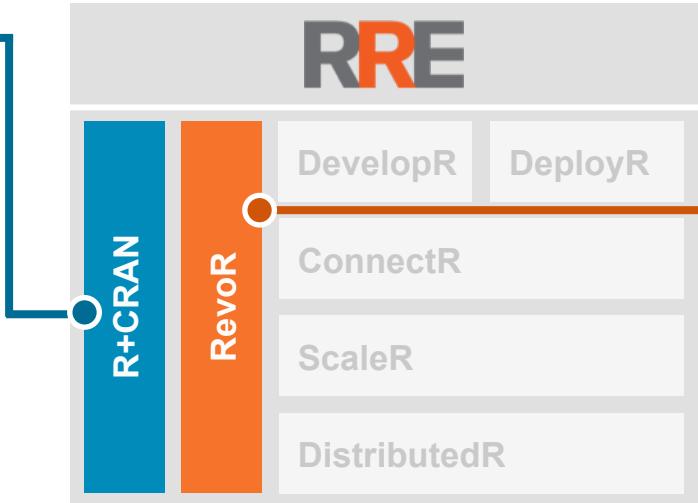


- Revolution R Enterprise includes all of the components you need for:
 - Enterprise readiness
 - High performance analytics
 - Multi-platform architecture support
 - Data source integration
 - Development tools
 - Deployment tools

The Platform Step by Step: R Capabilities

R+CRAN

- Open source R interpreter
 - **UPDATED** R 3.0.2
- Freely-available R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages



RevoR

- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math

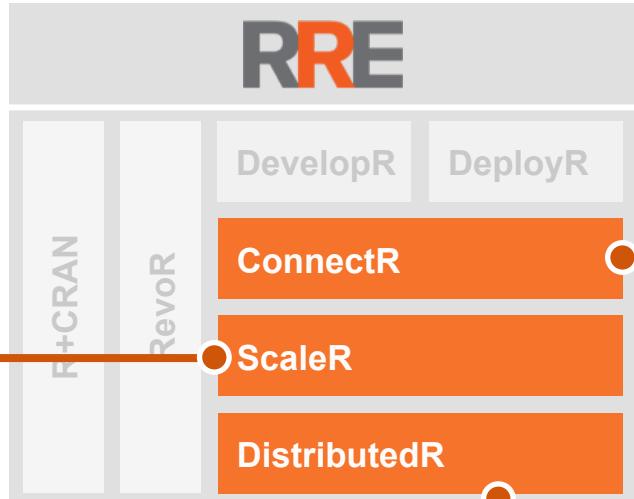
Available On:

- Platform™ LSF™ Linux®
- Microsoft® HPC Clusters
- Microsoft Azure Burst
- Windows® & Linux Servers
- Windows & Linux Workstations
- Teradata® Database
- IBM® Netezza®
- IBM BigInsights™
- Cloudera Hadoop®
- Hortonworks Hadoop
- Intel® Hadoop

The Platform Step by Step: Parallelization & Data Sourcing

ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Correlation & covariance matrices
- Predictive Models – linear, logistic, GLM
- Machine learning
- Monte Carlo simulation
- **NEW** Tools for distributing customized algorithms across nodes



ConnectR

- High-speed data import/export

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS & HBase
- Teradata Database TPT
- ODBC (incl. Vertica, Oracle, Pivotal, Aster, SybaseIQ, DB2, MySQL)

DistributedR

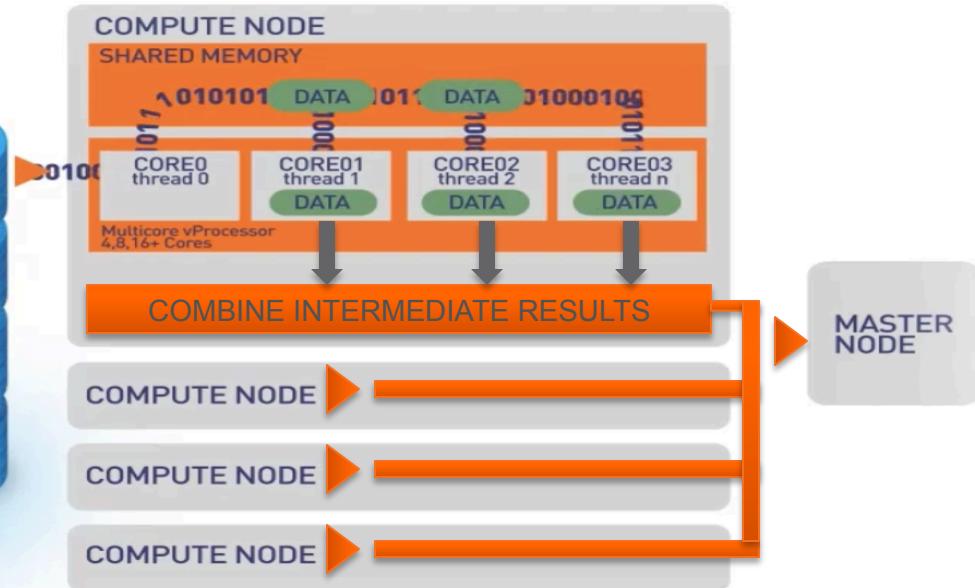
- Distributed computing framework
- Delivers portability across platforms

DistributedR available on:

- Windows Servers
- Red Hat and **NEW** SuSE Linux Servers
- IBM Platform LSF Linux Clusters
- Microsoft HPC Clusters
- Microsoft Azure Burst
- **NEW** Teradata Database
- **NEW** Cloudera Hadoop
- **NEW** Hortonworks Hadoop



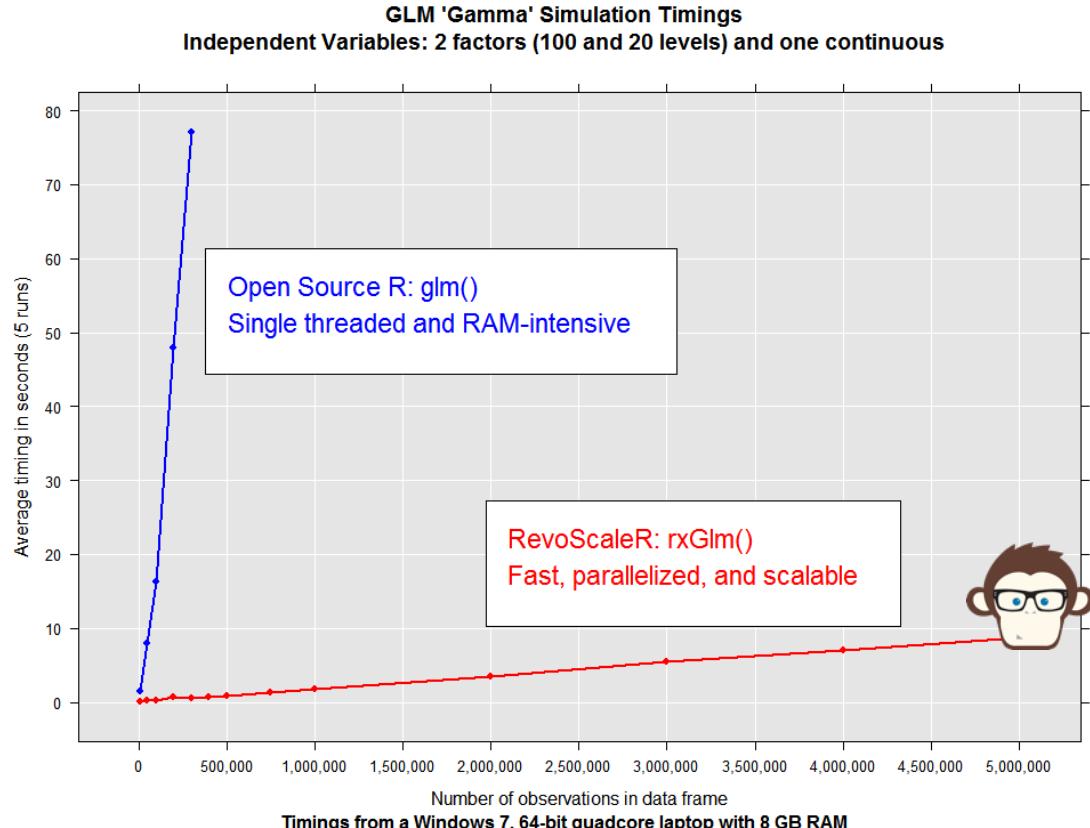
Powering Next Generation Analytics

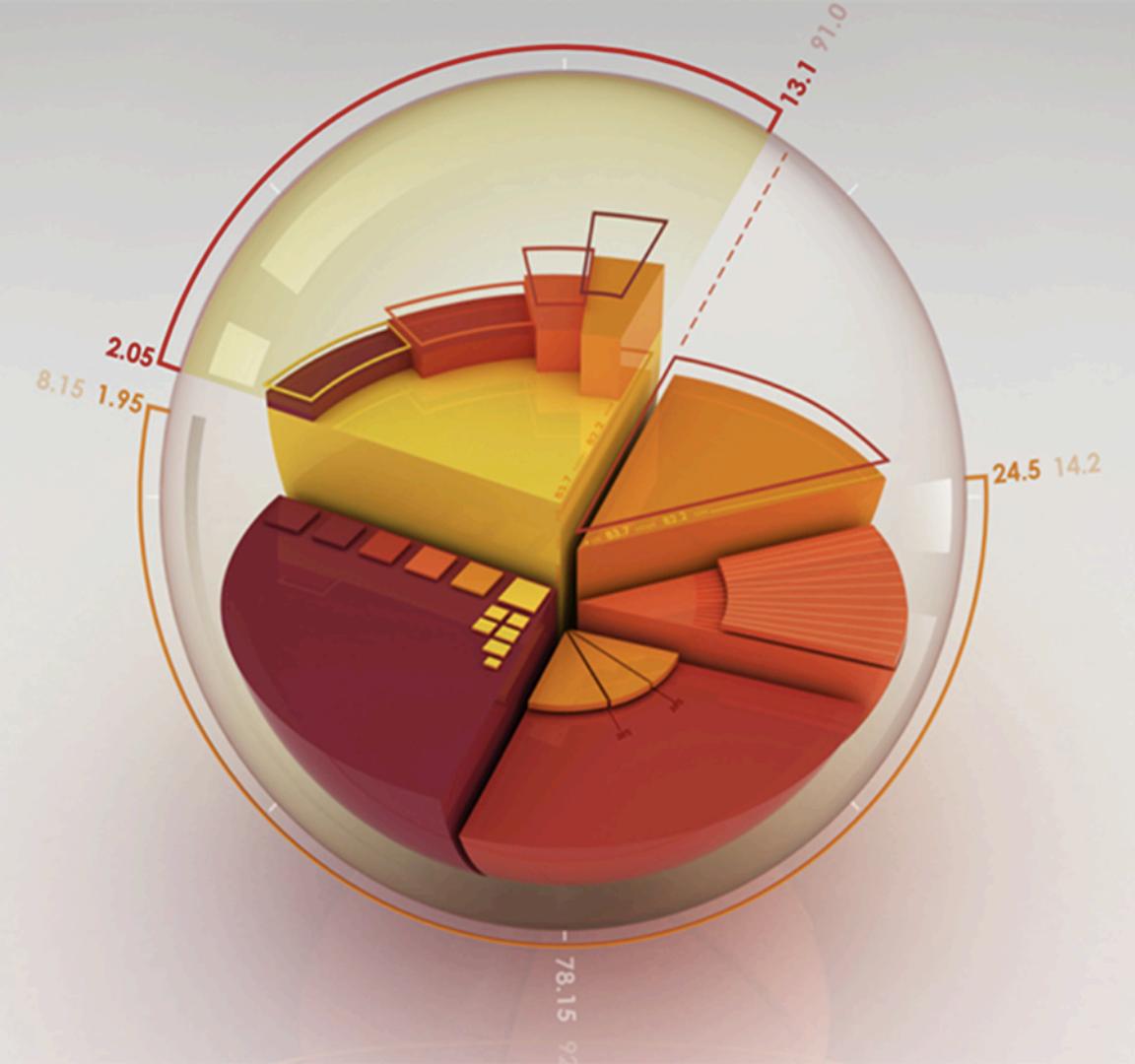


Eliminates Performance and Capacity Limits of Open Source R and Legacy SAS



- Unique PEMAs: Parallel, external-memory algorithms
- High-performance, scalable replacements for R/SAS analytic functions
- Parallel/distributed processing eliminates CPU bottleneck
- Data streaming eliminates memory size limitations
- Scales linearly with data size and compute capacity
- Works with in-memory and disk-based architectures







USING BIG
DATA
PLATFORMS

Bringing R to Big Data Architectures



Hadoop

Servers & Clusters

Data Warehouses

✓ **cloudera**

✓ **Hortonworks**

intel

IBM InfoSphere BigInsights
Bring the power of Hadoop to the enterprise.


✓ **{Platform Computing** IBM

✓ **Windows Server**

✓  **redhat.**

✓  **openSUSE**

✓ **TERADATA**

✓  **NETEZZA**

✓ Includes support for full suite of ScaleR algorithms on platform

Write Once, Deploy Anywhere



Teradata Database
Version 14.10



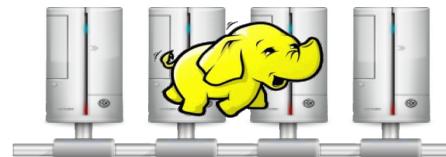
Microsoft
& Linux
Servers



Workstations

Write Once
Deploy Anywhere

Cloudera & Hortonworks
Hadoop



Server Clusters



Write Once → Deploy Anywhere

Set the desired compute context for code execution.....

**Local System
(default)**

→ rxSetComputeContext("local") # DEFAULT!!



→ rxSetComputeContext(RxLsfCluster(<data, server environment arguments>))



→ rxSetComputeContext(RxHpcServer(<data, server environment arguments>))



→ rxSetComputeContext(RxAzureBurst(<data, server environment arguments>))



→ rxSetComputeContext(RxHadoopMR(<data, server environment arguments>))

TERADATA

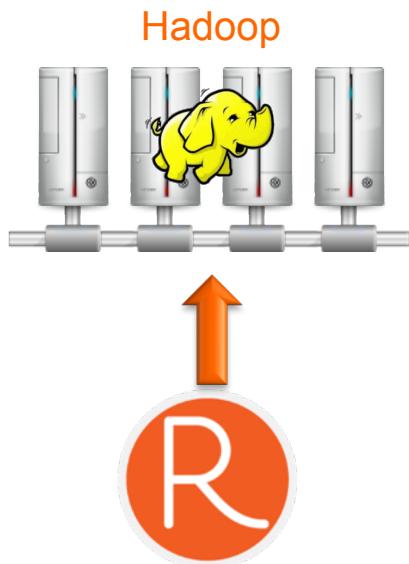
→ rxSetComputeContext(RxTeradata(<data, server environment arguments>))

Same code to be run anywhere

```
{  
  # Summarize and calculate descriptive statistics from the data airDS data set  
  adsSummary <- rxSummary(~ArrDelay+CRSDepTime+DayOfWeek, data = airDS)  
  
  # Fit Linear Model  
  arrDelayLm1 <- rxLinMod(ArrDelay ~ DayOfWeek, data = airDS); summary(arrDelayLm1)
```



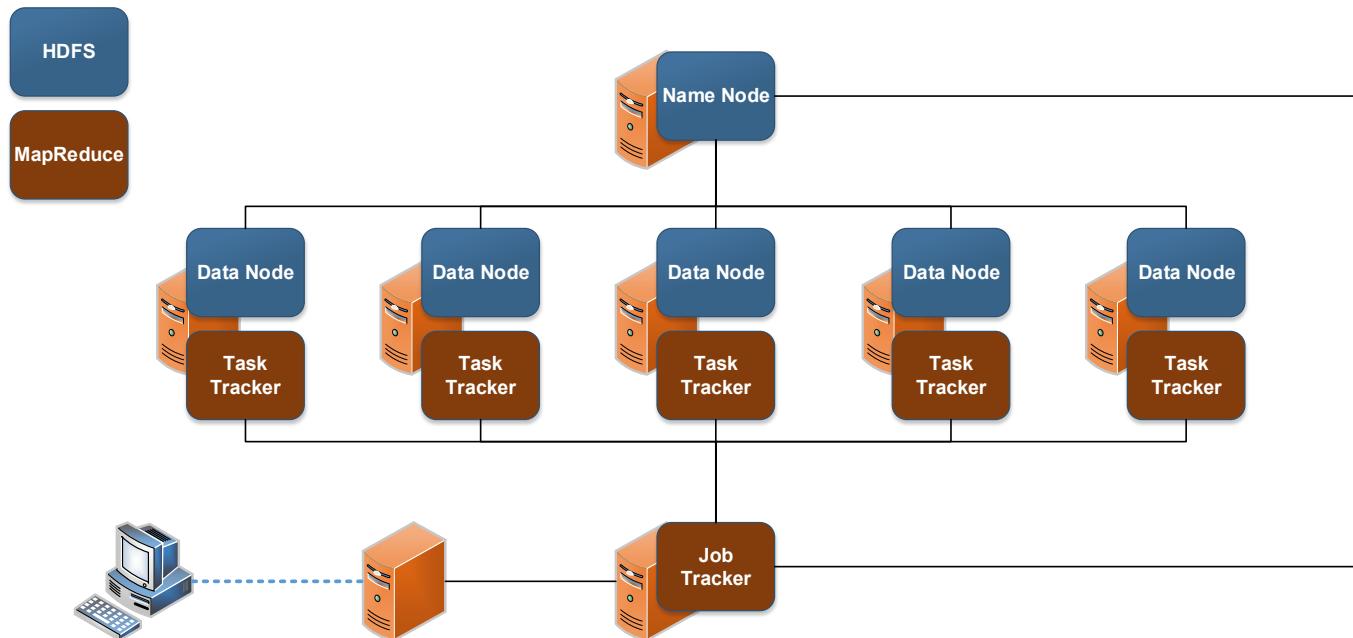
A Simple Goal: Hadoop As An R Engine.



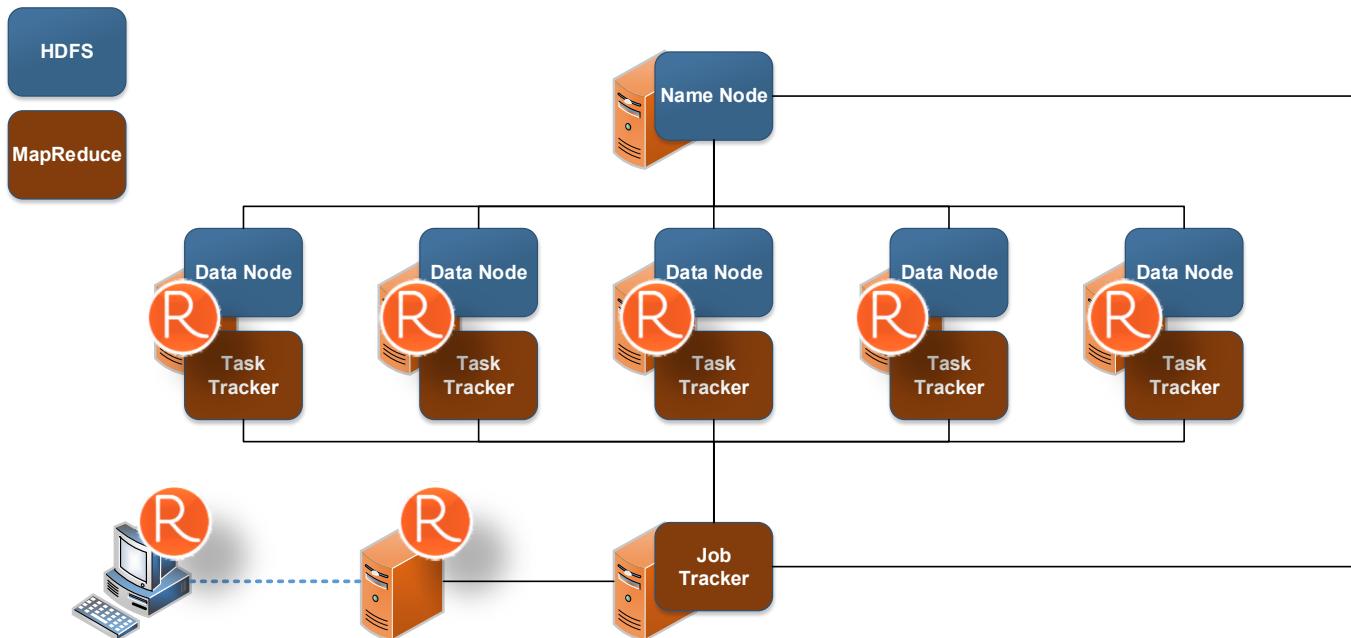
- Run Revolution R Enterprise code In Hadoop without change
- Provide RRE ScaleR Pre-Parallelized Algorithms
- Eliminate:
 - The need to “Think in MapReduce”
 - The need for a separate compute cluster
 - Data movement

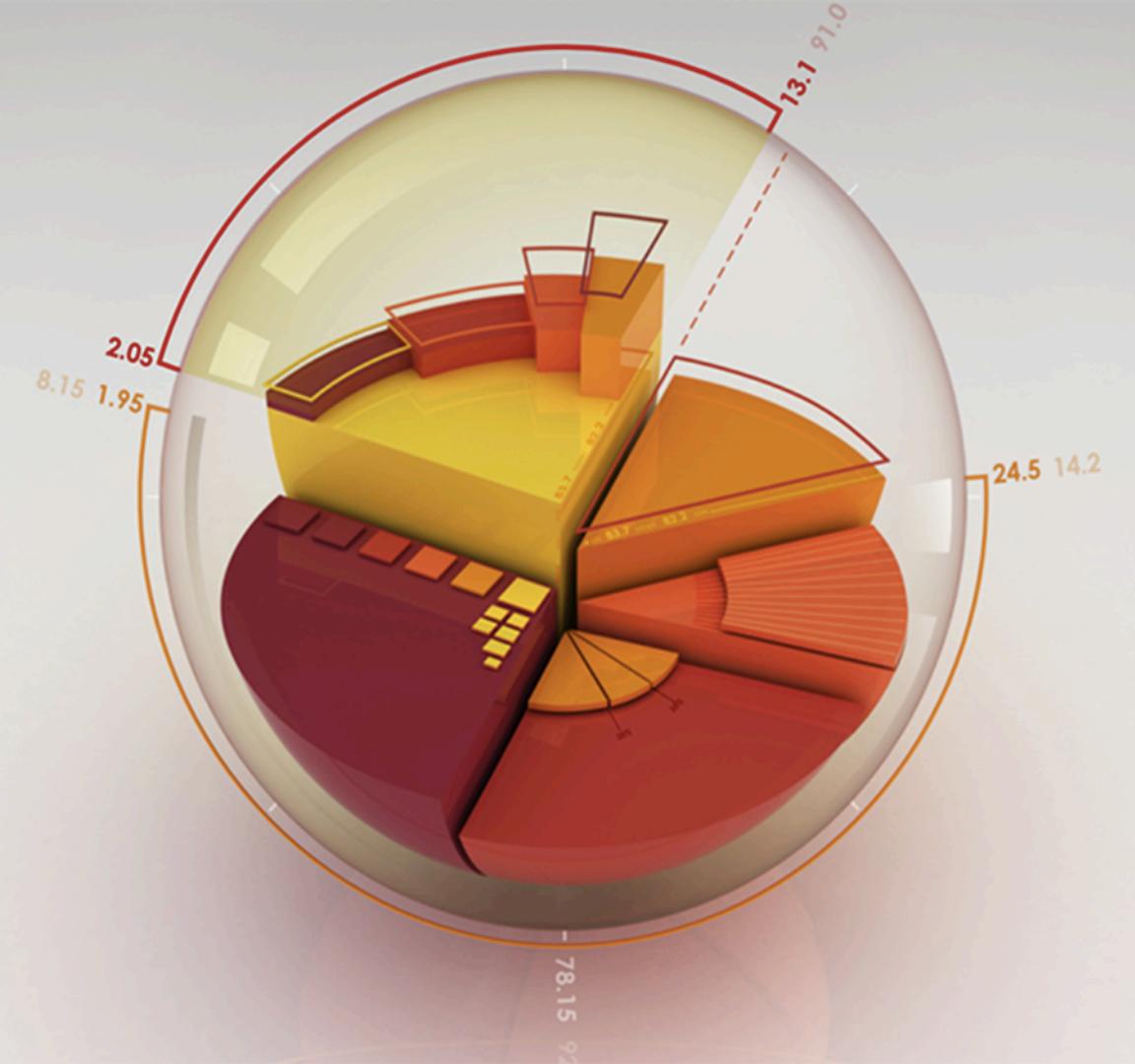


RRE in Hadoop



RRE in Hadoop







DEPLOYMENT THE LAST MILE PROBLEM



The Platform Step by Step: Tools & Deployment

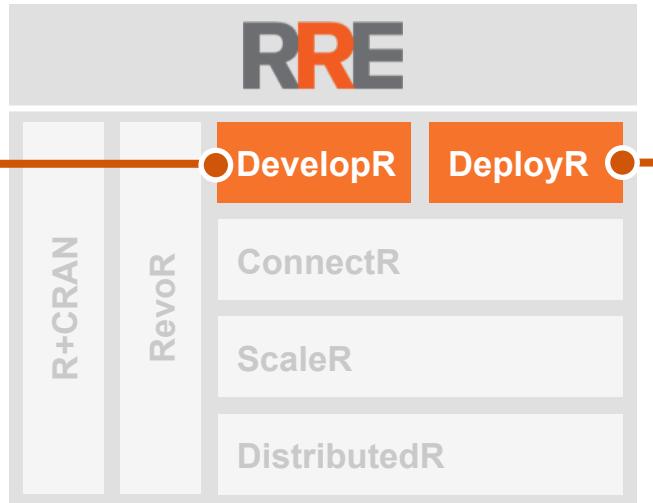
DevelopR

- Integrated development environment for R
- Visual 'step-into' debugger

Available on:

- Windows

Or use:



DeployR

- Web services software development kit for integration analytics via Java, JavaScript or .NET APIs
- Integrates R Into application infrastructures

Capabilities:

- Invokes R Scripts from web services calls
- RESTful interface for easy integration
- Works with web & mobile apps, leading BI & Visualization tools and business rules engines





Custom Integration with Web Services API

RRE DeployR makes R accessible

```
1 ## Analysis!
2 ##
3 # transaction level data
4 myReportProgress<-0
5
6
7 ## Calculate correlations on transaction-based data
8 catNames <- names(rxGetInfoXdf(file=currentTransSubset, getVa
9 catNames <- catNames[!(catNames %in% c("transaction_id", "sta
10 corFormula <- paste("correlation", "transaction_id", "type"))
11 corFormula <- as.formula(corFormula)
12 corFormula <- as.formula(corFormula)
13
14 # the actual correlation calculation
15 t2 <- rxCorCor(formula=corFormula, data=currentTransSubset,
16 type = "Cor", reportProgress=myReportProgress)[["CovCor"]]
17
18 # remove empty rows and columns
19 diag(t2) <- NA
20 for (iCol in ncol(t2)-1) {
21   if (!all(is.na(t2[,iCol]))) {
22     t2 <- t2[-iCol, iCol]
23   }
24 }
25 diag(t2) <- 1
26
27 ## Cluster and plot cluster dendrogram
```

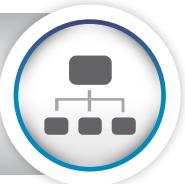
R / Statistical Modeling Expert



RRE DeployR



Application Developer



- Seamless
 - Bring the power of R to any web enabled application
- Simple
 - Web Services API leverages application development frameworks including JS, Java, .NET
- Scalable
 - Robustly scale user and compute workloads
- Secure
 - Manage enterprise security with LDAP & SSO

Data Analysis



Business Intelligence



Mobile Web Apps



Cloud / SaaS



App Integration

QlikView - [What's New in QV9]

File Edit View Selections Layout Settings Bookmarks Reports Tools Object Window Help

Introduction Background Dashboard Budget Customers Sales Sales Rep Transactions

Search

Current Selections

Fields **CURRENCY** EUR Values

Year **2008**

Sales EUR (K) Margin EUR (K) Ma

2008	58,279	24,204
2007	55,865	21,913
Variance	4.3%	10.5%

Regional Scorecard

Region	Ranking	Sales Trends 2007 - 2008	Sales 2008
Total			58,279
NORDIC	23	22,000	23,000
USA	24	16,394	17,404
JAPAN	25	7,386	7,395
SPAIN	26	2,359	2,100
GERMANY	27	2,100	2,100

Region

GERMANY	JAPAN	NORDIC
UK	USA	SPAIN

2008 vs 2007 Sales = 104%

0% 25% 50% 75% 100% 125% 150%

2008 Margin 42%

0% 10% 20% 30% 40% 50%

2008 Actual vs Budget 97%

0% 25% 50% 75% 100% 125% 150%

View Finance

Data refreshed on 04/04/2009

For Help, press F1

4/4/2009 4:52:38 PM D: 1/3 F: 42346/128404

Sales Data - Microsoft Excel

Sheet Insert Page Layout Formulas Data Review View

Tables PivotTable PivotChart Charts

H18

Total Sales by Store

Store	Sales
Store 1	\$ 313,765
Store 2	\$ 107,160
Store 3	\$ 351,751
Store 4	\$ 131,047
Store 5	\$ 252,136
Store 6	\$ 167,462
Store 7	\$ 210,073
Store 8	\$ 308,099
Store 9	\$ 97,492
Store 10	\$ 393,484
Store 11	\$ 396,891
Store 12	\$ 151,168
Store 13	\$ 251,390
Store 14	\$ 392,776
Store 15	\$ 259,654
Store 16	\$ 225,184
Store 17	\$ 335,785
Total	\$ 4,162,346

Total Sales by Region

Region	Sales
West	\$ 1,718,238
South	\$ 534,389
Midwest	\$ 1,009,268
East	\$ 900,431

TOTAL SALES BY CATEGORY

Total Sales by Category

Category	Sales
Automotive	\$ 86,285
Gardening	\$ 52,048
Electronics	\$ 83,026
Jewelry	\$ 93,035
Sporting	\$ 50,016
Books	\$ 42,247
Games	\$ 18,420



Business Analysts: Alteryx

Alteryx Designer x64

File Edit View Tools Window Help

Run Module Ctrl+R Refresh Config F5

Schedule Module... View Schedules...

Install License... Manage License...

User Settings

- Edit User Settings...
- Save Layout & Settings on Exit
- Save Layout & Settings Now
- Restore Defaults

Module Runtime Path: C:\Program Files\Alteryx\Designer\dlls\

Type: Module Analytic App Macro: Standard Macro

Canvas Options

Layout Direction: Horizontal Annotations: Show Connection Progress: Show Only When Running

Constraints:

Type	Name	Value
Engine	TempFilePath	C:\Users\SoftpedialEditor\AppData\Local\Temp\Alteryx\
Engine	Version	8.5.3.39602
Engine	ModuleDirectory	C:\Program Files\Alteryx\Designer\dlls\
Engine	GuiInteraction	True

Output: 0 Errors 0 Conv Errors 0 Warnings 0 Messages 0 Files All

Multiple Alteryx Module: Iterative Supply and Demand Scenario

Iterative Supply and Demand Scenario

Remaining Supply (should be empty)

Optimal Search Distance

Check that Supply matches Assigned Demand

Assigned Pairs

Σ

Σ #1

Σ #2

sample_demand.yxdb

sample_supply.yxdb

How do I view the Macro tool?

Right Click on the Macro tool to open the underlying Iterative Macro Module.

Check out the Design Settings to see how it works!

Expand this container to display tool descriptions

Use the input tool to bring in the point spatial objects for Supply and Demand nodes.

Create a unique record ID for Supply and Demand nodes. Starting each at a different number ensures they are unique across node types.

The Append tool creates a cartesian join between each pair of Supply and Demand nodes.

The distance tool calculates the distance between each pair of Supply and Demand nodes.

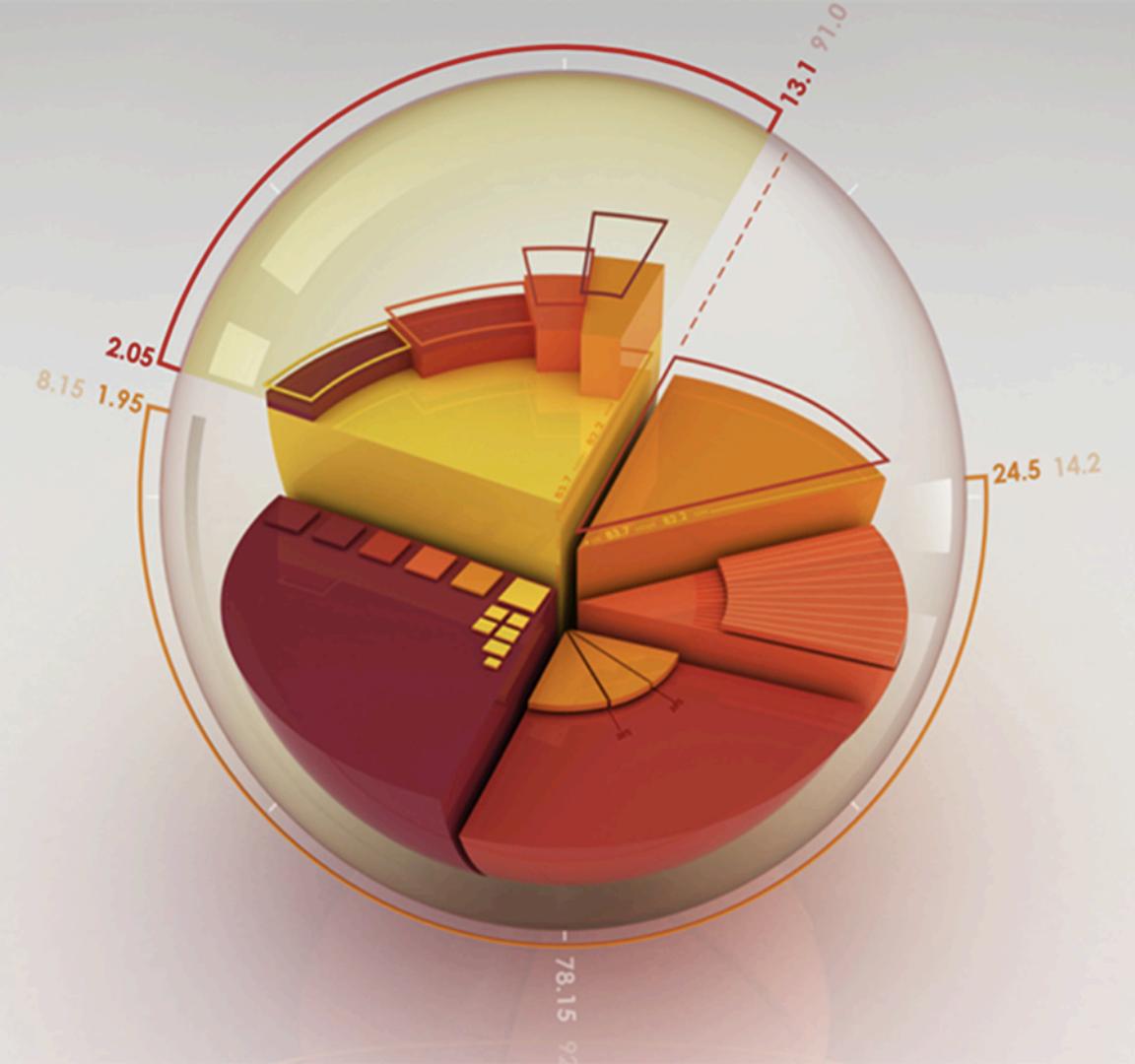
The records are sorted based on their distance in miles in ascending order.

The select tool is used to remove unnecessary records.

The Iterative and Demand tool iterates until all the supply is assigned to available demand.

A series of tools are used to calculate the maximum distance between assigned supply and demand locations.

The Assign tool identifies the assigned Supply and Demand nodes.





With Thanks

- The R Core Team
- R developers (5000 packages on CRAN!)
- The R community
- You!

David Smith @revodavid
david@revolutionanalytics.com

www.revolutionanalytics.com
1.855.GET.REVO
Twitter: @RevolutionR

Michael Helbraun
michael.helbraun@revolutionanalytics.com



Thank you.

www.revolutionanalytics.com

1.855.GET.REVO

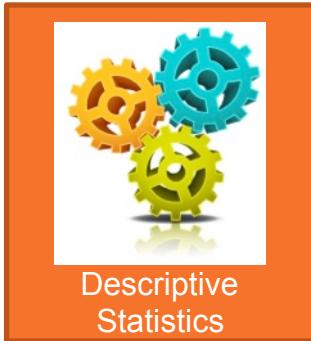
Twitter: @RevolutionR



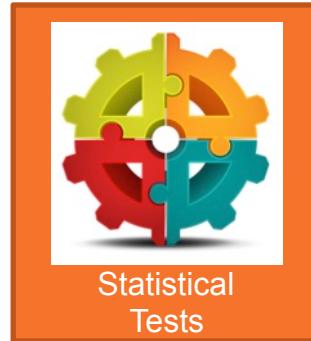
High Performance Big Data Analytics with Revolution R Enterprise ScaleR



R Data Step



Descriptive Statistics



Statistical Tests



Sampling



Predictive
Modeling



Data
Visualization



Machine
Learning



Simulation





Revolution R Enterprise ScaleR: High Performance Big Data Analytics

Data Prep, Distillation & Descriptive Analytics

R Data Step



- Data import – Delimited, Fixed, SAS, SPSS, ODBC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort
- Merge
- Split
- Aggregate by category (means, sums)

Descriptive Statistics



- Min / Max
- Mean
- Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests



- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling



- Subsample (observations & variables)
- Random Sampling



Revolution R Enterprise ScaleR (continued)

Statistical Modeling

Predictive Models



- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM)
 - All exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions including: cauchit, identity, log, logit, probit.
 - User defined distributions & link functions.
- Covariance Matrix
- Correlation Matrix
- Logistic Regression
- Classification & Regression Trees
- Residuals for all models

Data Visualization



- Histogram
- Line Plot
- Scatter Plot
- Lorenz Curve
- ROC Curves (actual data and predicted values)
- **NEW** Tree Visualization

Variable Selection



- Stepwise Regression
 - Linear
 - **NEW** logistic
 - **NEW** GLM
- Monte Carlo

Simulation



Machine Learning

Cluster Analysis



- K-Means

Classification



- Decision Trees
- **NEW** Decision Forests

Deployment



- Prediction (scoring)
- **NEW** PMML Export