REVOLUTION ANALYTICS

Revolution R Enterprise

Portland R User Group

November 13, 2013

David Smith @revodavid
Michael Helbraun

BIG DATA


DATA
SCIENCE

OPEN SOURCE R

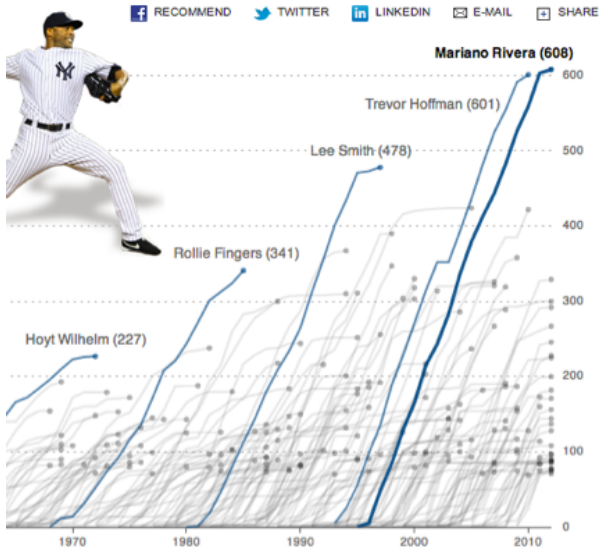
Innovate with R

- 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

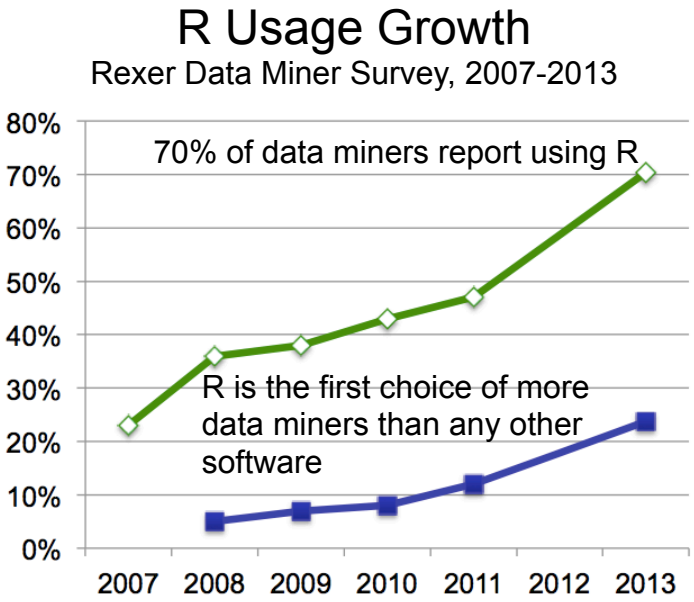
bit.ly/r-is-hot

WHITE PAPER



REVOLUTION ANALYTICS

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 ANALYTICS

Revolution R Enterprise

Power R for the Enterprise

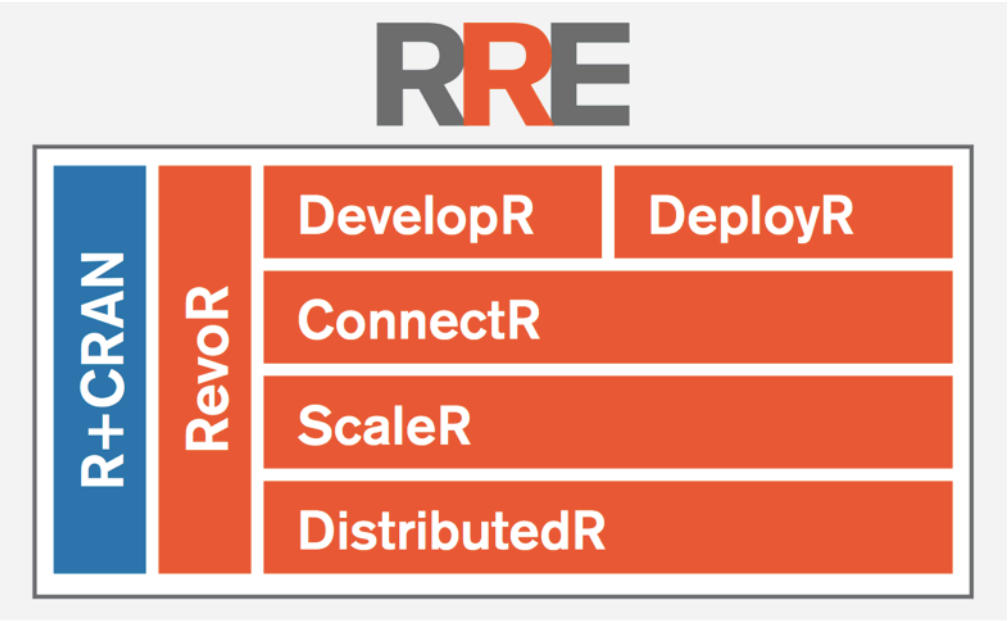
Supercharge R for Massive Data

Empower Platform Independence

Take Big Cost Out of Big Data

RRE is the **Big Data Big Analytics Platform**

REVOLUTION R ENTERPRISE

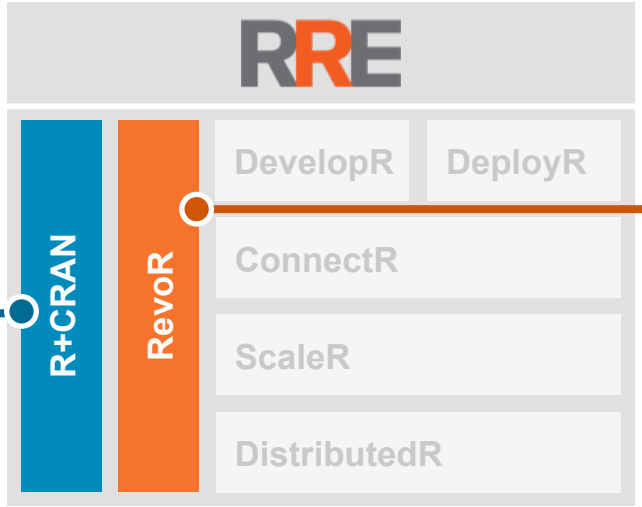


- Revolution R Enterprise includes all of the components you need for:
  - Enterprise readiness
  - High performance analytics
  - Multi-platform architecture support
  - Data source integration
  - Development tools
  - Deployment tools

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:

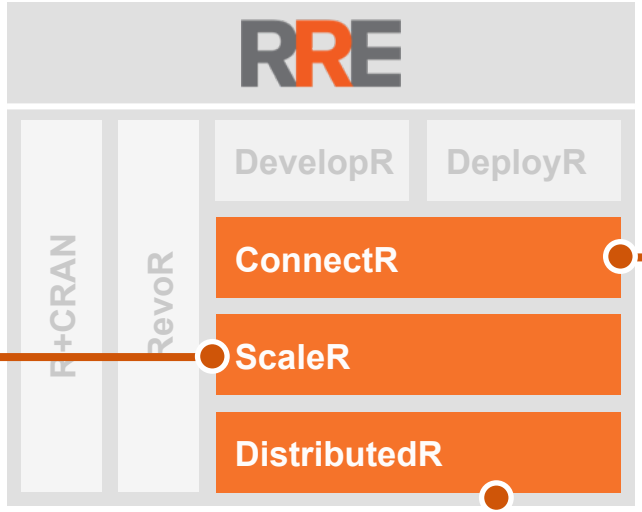
- Platform™ LSF™ Linux®
- Microsoft® HPC Clusters
- Microsoft Azure Burst
- Windows® & Linux Servers
- Windows & Linux Workstations
- Teradata® Database
- IBM® Netezza®
- IBM BigInsights™
- Cloudera Hadoop®
- Hortonworks Hadoop
- Intel® Hadoop

The Platform Step by Step: **Parallelization & Data Sourcing**

REVOLUTION R ENTERPRISE

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

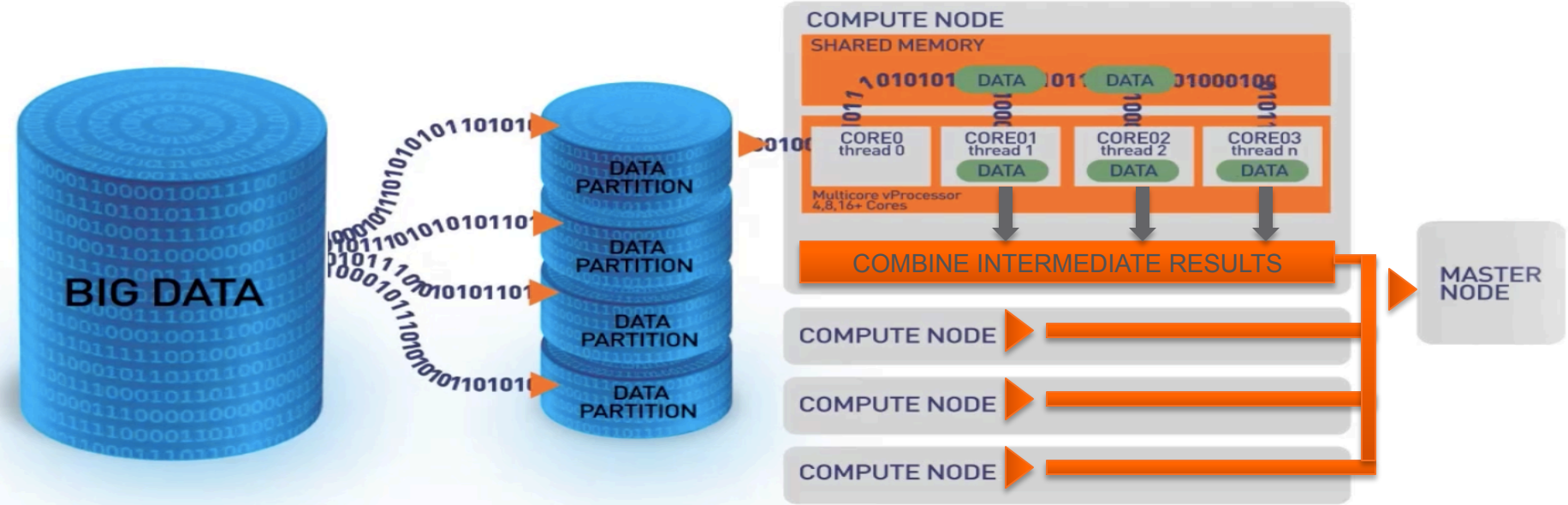
- Distributed computing framework
- Delivers portability across platforms

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

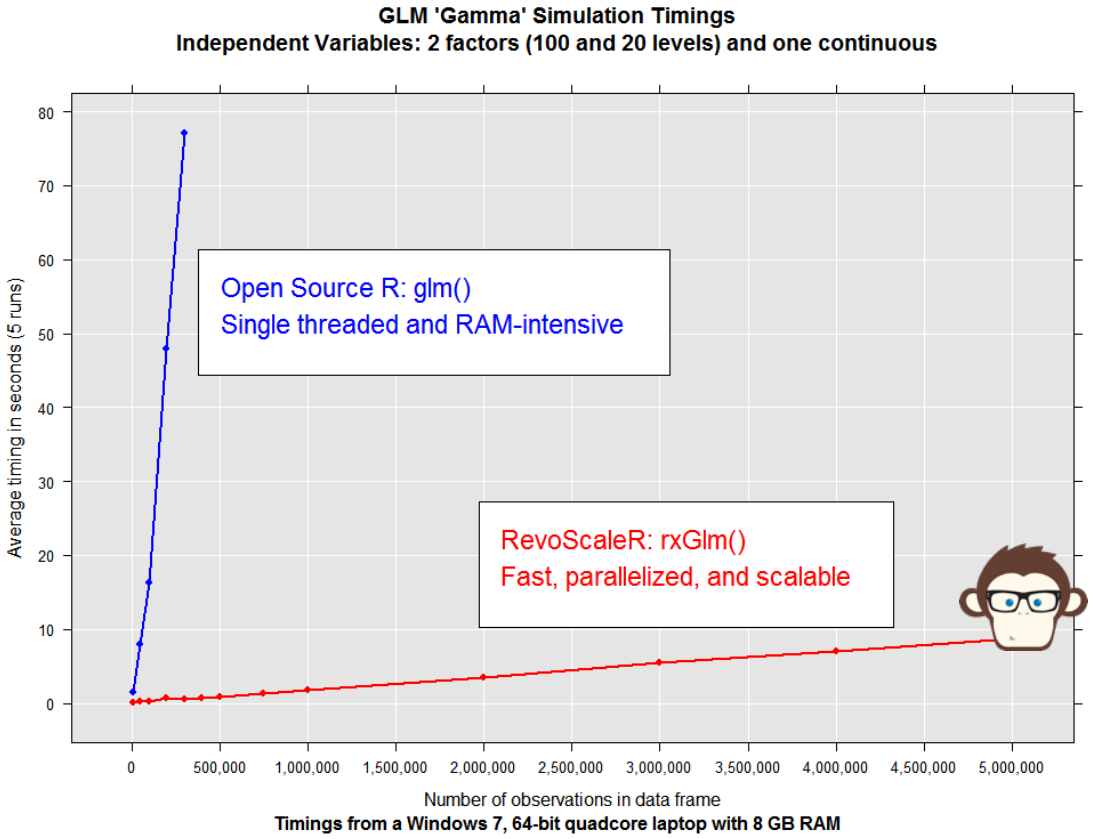
REVOLUTION ANALYTICS

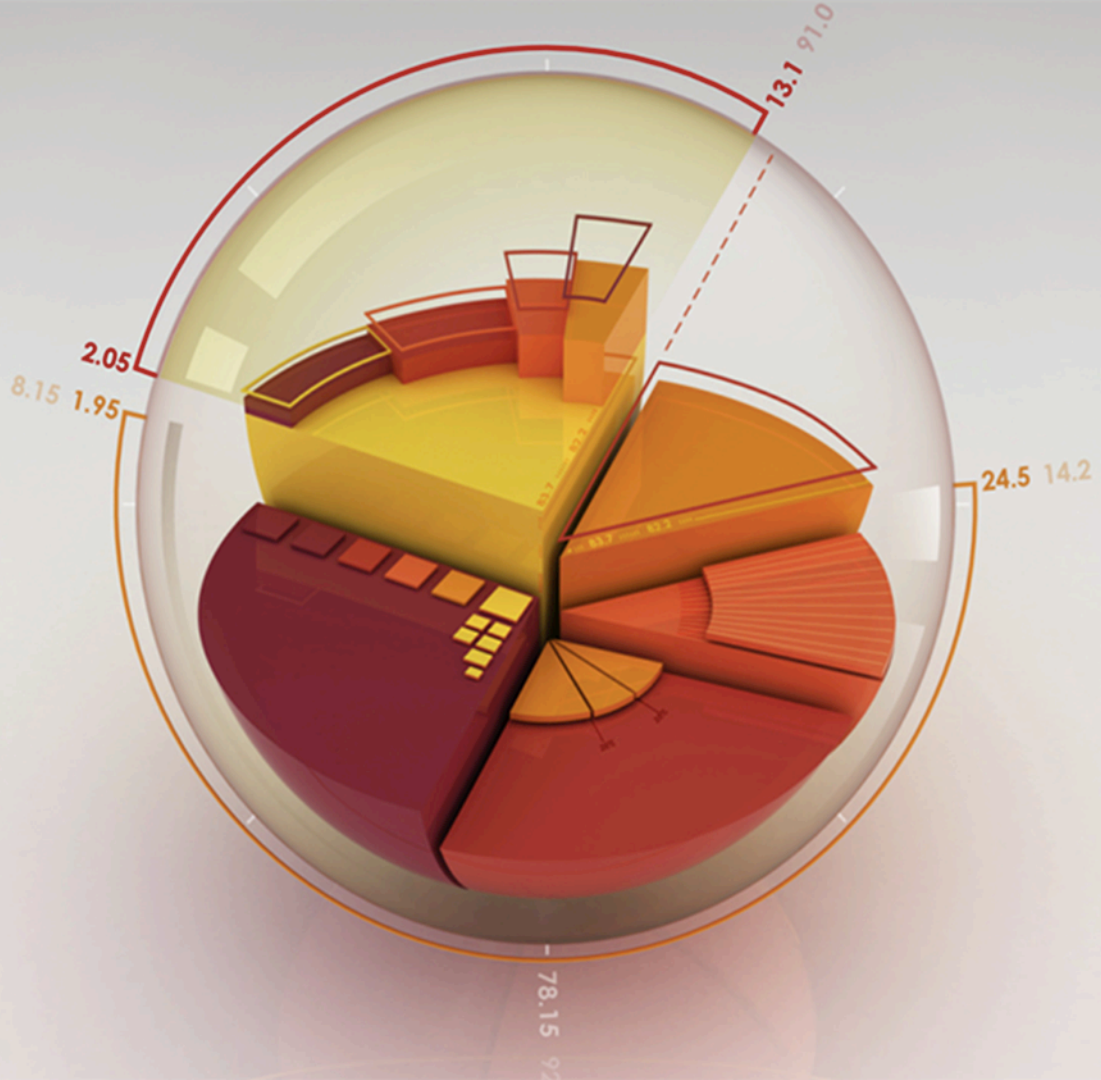
Powering Next Generation Analytics



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





DEMO

USING BIG DATA PLATFORMS

Bringing R to Big Data Architectures

Hadoop

- ✓ cloudera
- ✓ Hortonworks
- intel
- IBM InfoSphere BigInsights — Bring the power of Hadoop to the enterprise.

Servers & Clusters

- ✓ Platform Computing (IBM)
- ✓ Windows Server
- ✓ redhat.
- ✓ openSUSE

Data Warehouses

- ✓ TERADATA
- NETEZZA

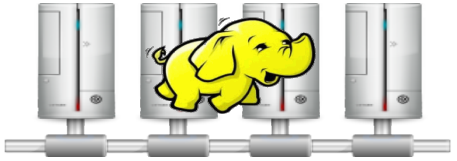
✓ Includes support for full suite of ScaleR algorithms on platform

REVOLUTION ANALYTICS

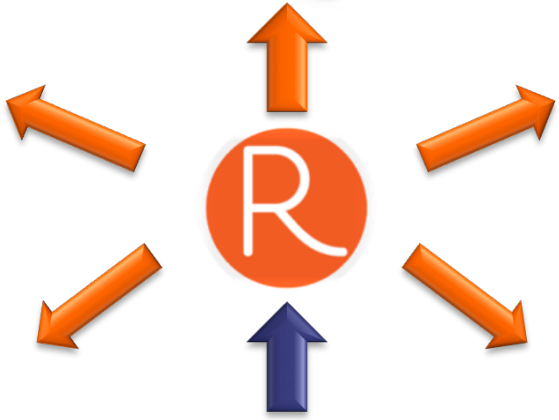
✓ Write Once, Deploy Anywhere

Teradata Database
Version 14.10

Cloudera & Hortonworks
Hadoop



Microsoft
& Linux
Servers



Workstations

Write Once
Deploy Anywhere

Server Clusters

REVOLUTION ANALYTICS

Write Once → Deploy Anywhere

Set the desired compute context for code execution…..

Local System (default) →
```
rxSetComputeContext("local") # DEFAULT!!
```

redhat. Platform Computing →
```
rxSetComputeContext(RxLsfCluster(<data, server environment arguments>))
```

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

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

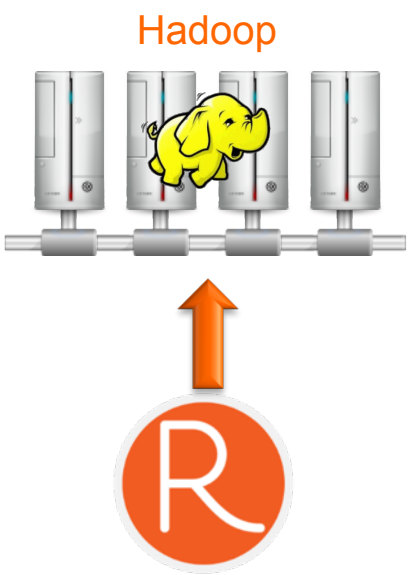
hadoop cloudera Hortonworks →
```
rxSetComputeContext(RxHadoopMR(<data, server environment arguments>))
```

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

Same code to be run anywhere …..

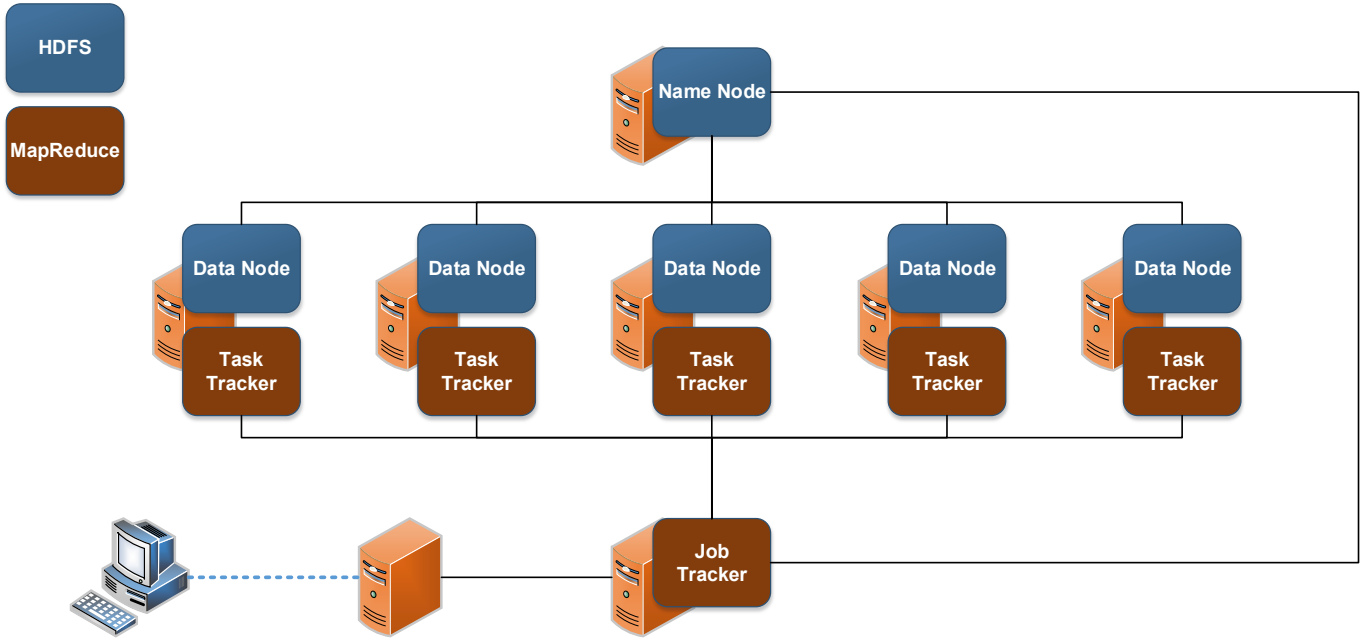
```
# Summarize and calculate descriptive statistics from the data airDS data set
adsSummary <- rxSummary(~ArrDelay+CRSDepTime+DayOfWeek, data = airDS)

# Fit Linear Model
arrDelayLm1 <- rxLinMod(ArrDelay ~ DayOfWeek, data = airDS); summary(arrDelayLm1)
```

A Simple Goal: Hadoop As An R Engine.

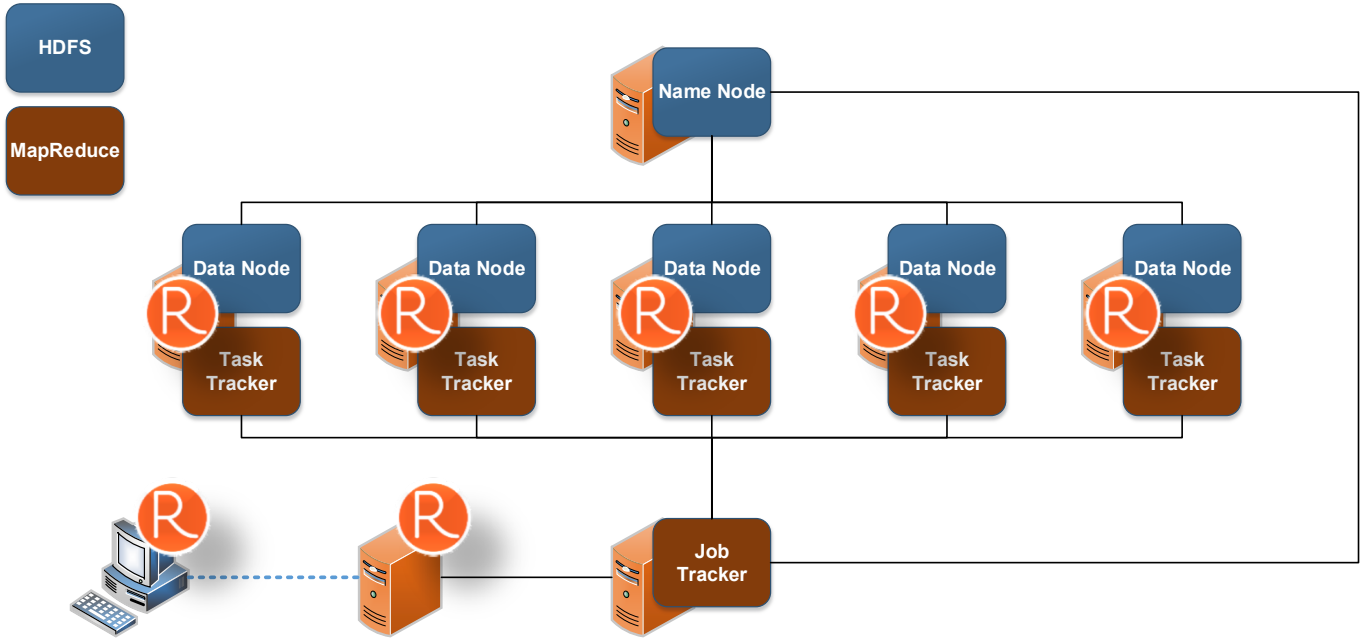


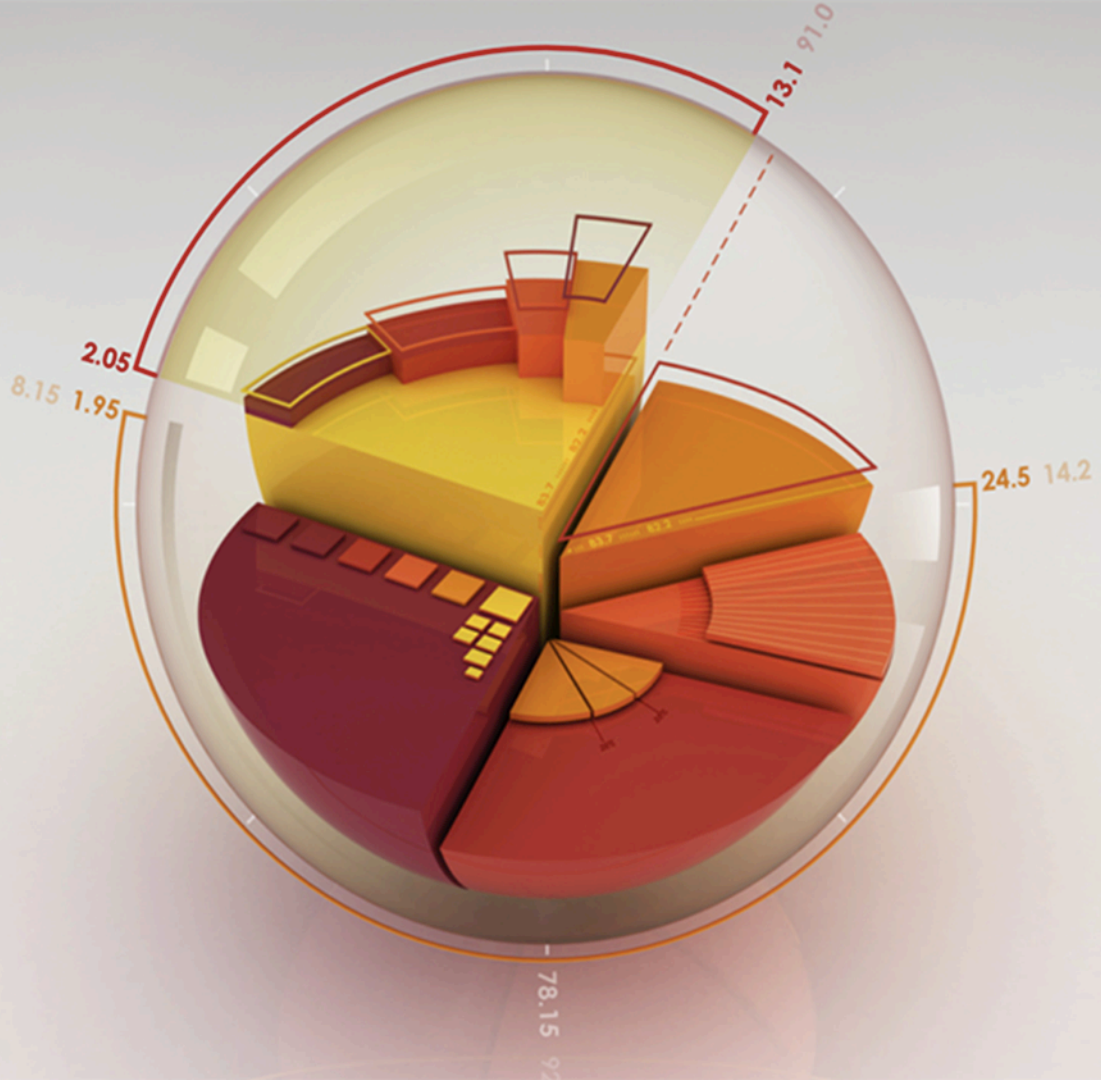
- Run Revolution R Enterprise code In Hadoop without change
- Provide RRE ScaleR Pre-Parallelized Algorithms
- Eliminate:
  - The need to “Think in MapReduce”
  - The need for a separate compute cluster
  - Data movement

RRE in Hadoop



RRE in Hadoop





DEMO

DEPLOYMENT

THE LAST MILE PROBLEM

The Platform Step by Step: **Tools & Deployment**

DevelopR

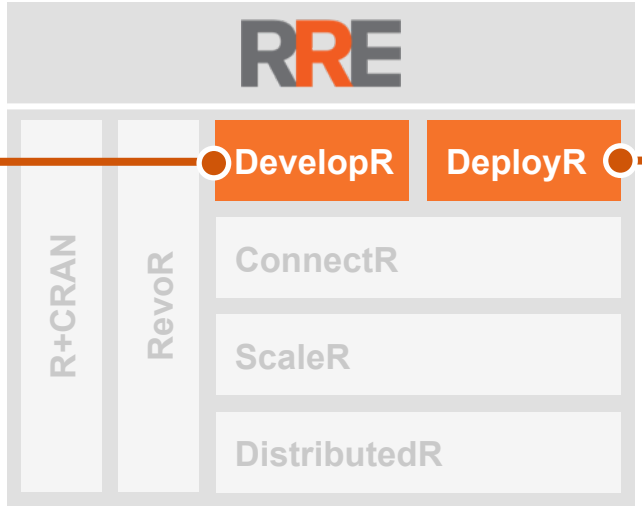
- Integrated development environment for R
- Visual 'step-into' debugger

Available on:

- Windows

Or use:

RStudio



DeployR

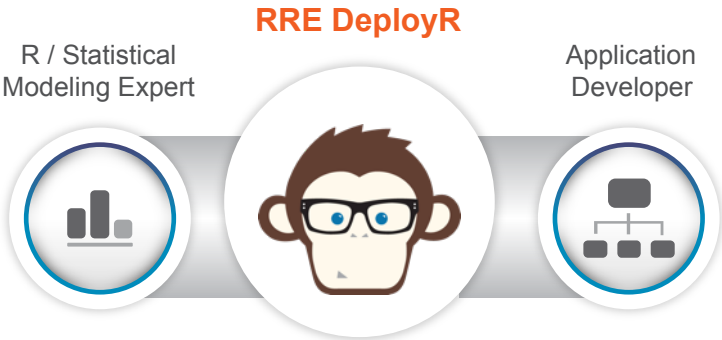
- Web services software development kit for integration analytics via Java, JavaScript or .NET APIs
- Integrates R Into application infrastructures

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

```
###
### Analysis!
###
# transaction level data
myReportProgress<-0

## Calculate correlations on transaction-based data
catNames <- names(rxGetInfoXdf(file=currentTransSubset, getVa
catNames <- catNames[!(catNames %in% c("transaction_id", "sta
corFormula <- paste(catNames, collapse="+")
corFormula <- paste("~", corFormula)
corFormula <- as.formula(corFormula)

# the actual correlation calculation
t2 <- rxCovCor(formula=corFormula, data=currentTransSubset,
type = "Cor", reportProgress=myReportProgress)[["CovCor"]]

# remove empty rows and columns
diag(t2) <- NA
for (iCol in ncol(t2):1) {
  if (all(is.na(t2[,iCol]))) {
    t2 <- t2[-iCol,-iCol]
  }
}
diag(t2) <- 1

## cluster and plot cluster dendogram
```



- 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

tableau

Business Intelligence

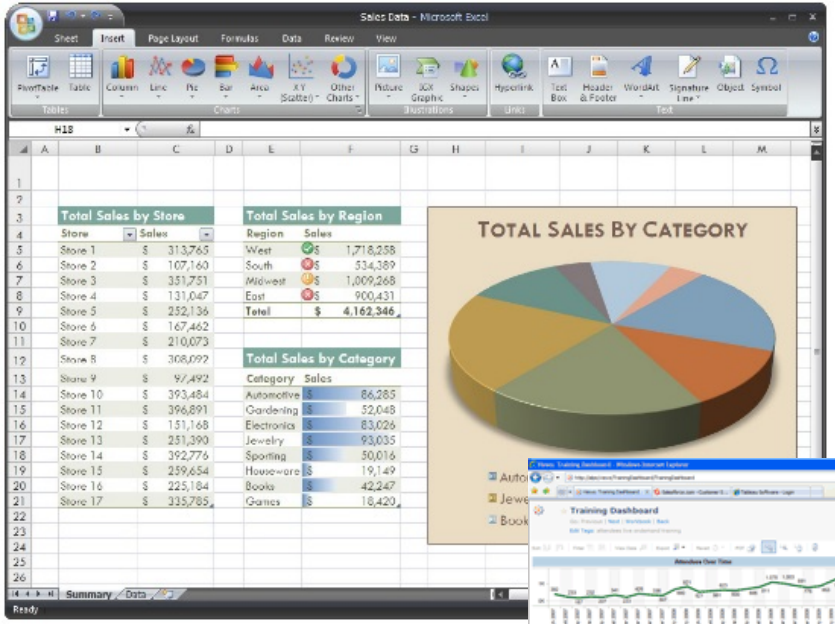
Mobile Web Apps

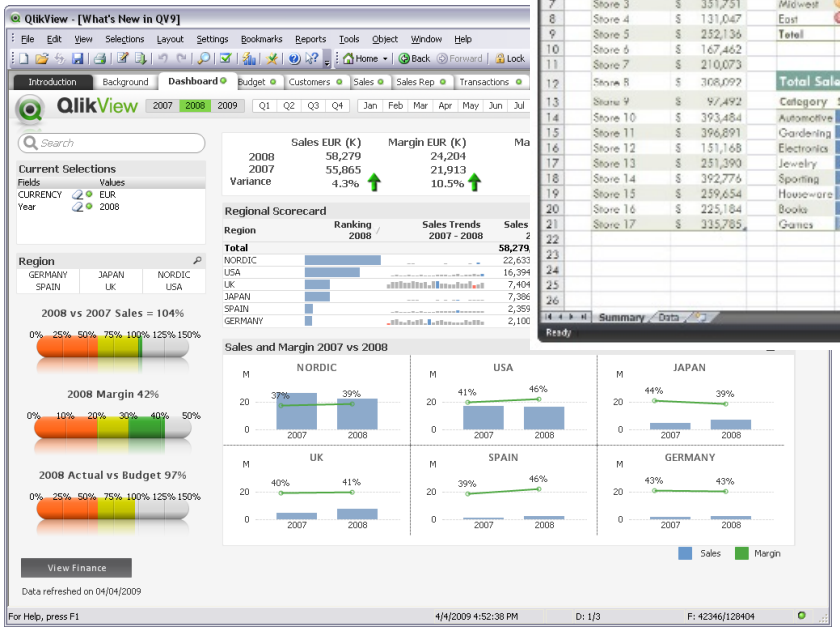
Cloud / SaaS

amazon.com

REVOLUTION ANALYTICS

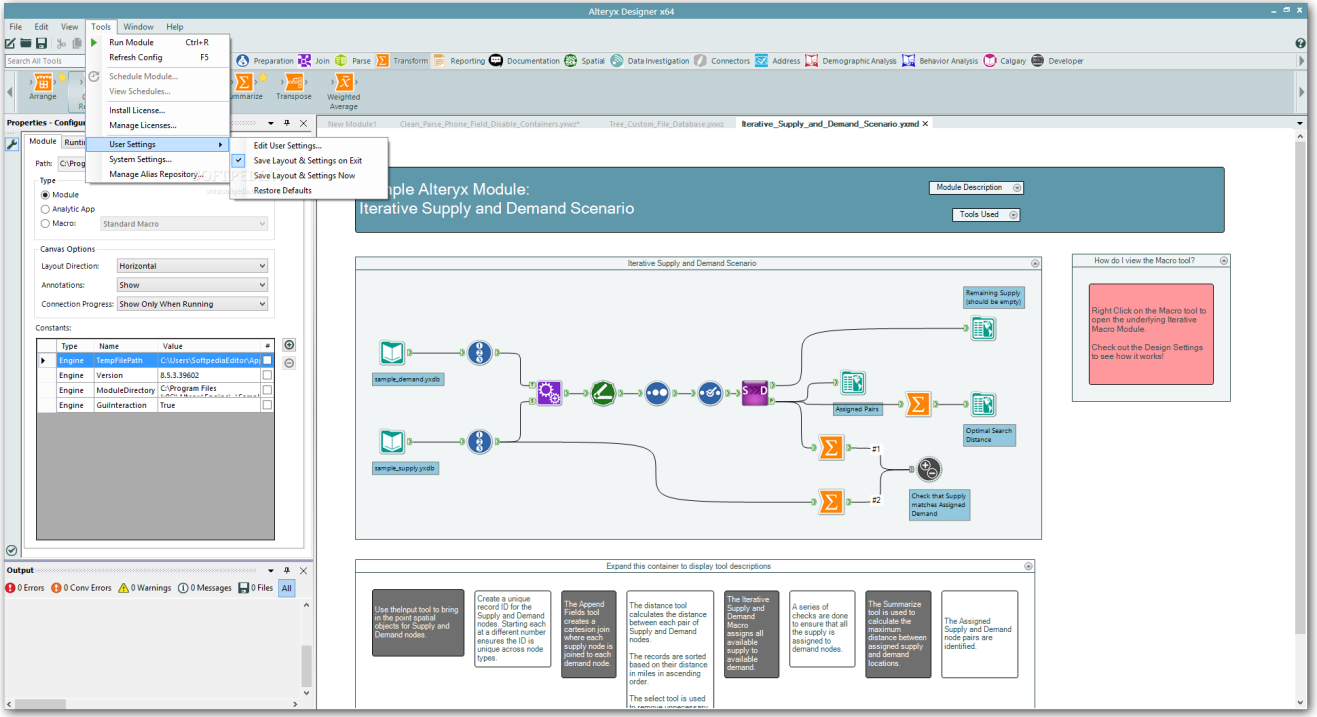
App Integration

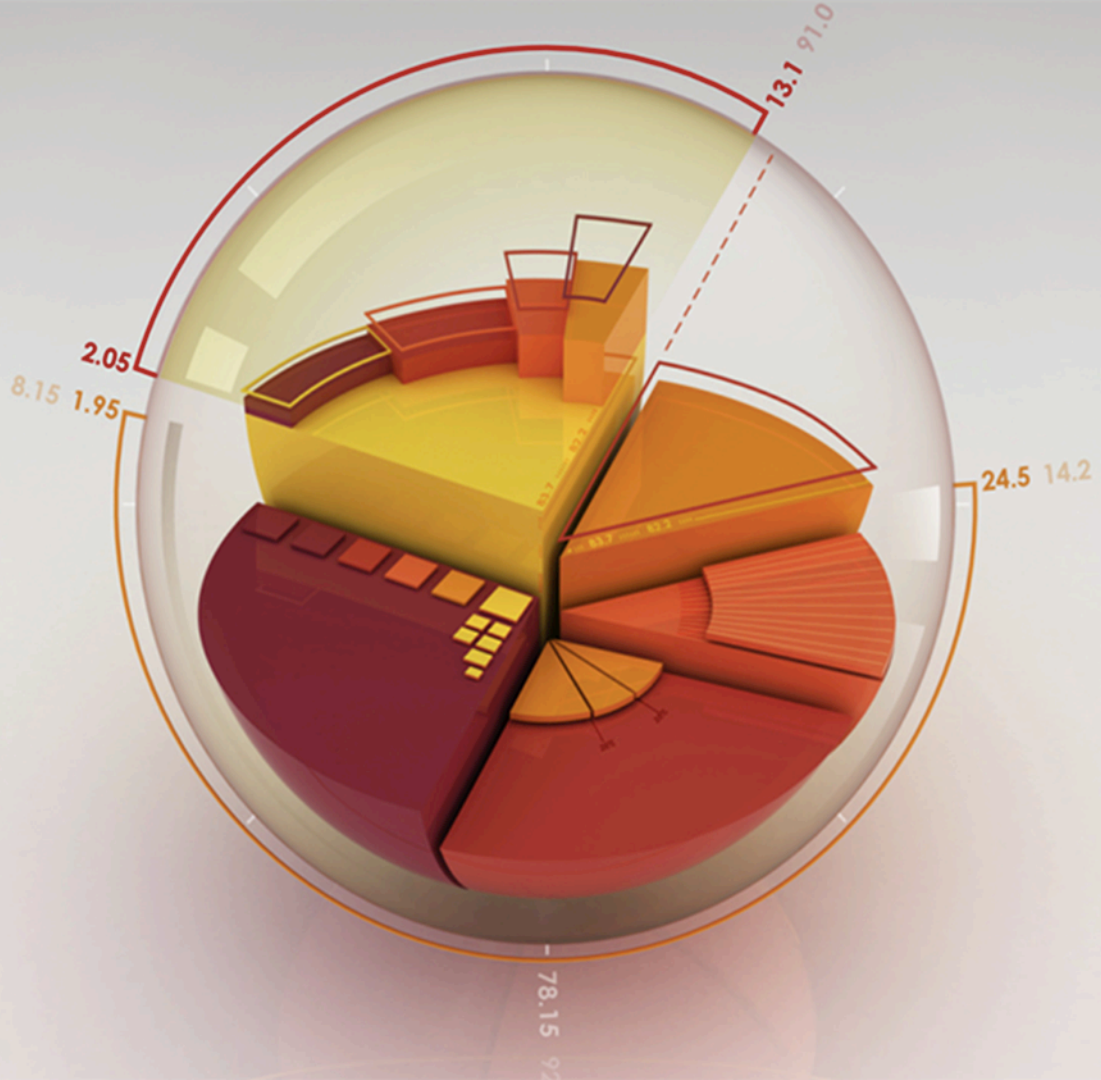




REVOLUTION ANALYTICS

Business Analysts: Alteryx





DEMO

With Thanks

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

David Smith @revodavid
david@revolutionanalytics.com

Michael Helbraun
michael.helbraun@revolutionanalytics.com

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

REVOLUTION ANALYTICS

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 ANALYTICS

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 ANALYTICS

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

Simulation

- Monte Carlo

Machine Learning

Cluster Analysis

- K-Means

Classification

- Decision Trees
- **NEW** Decision Forests

Deployment

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