Presentation for the NJ Hadoop User Group: R and Hadoop Working Together

Presented by:
Steve Belcher Ph.D.

Revolution Analytics
@RevolutionR
Agenda

- Introduction
- Who is Revolution Analytics?
- Quick Intro to R and Hadoop
- R and Hadoop together: Options
- Case Studies
- Revolution R Enterprise demonstration
- Q & A
INTRODUCTION
WHO IS REVOLUTION ANALYTICS?
Who we are
Leading provider of commercial analytics platform based on open source R statistical computing language

Our Software Delivers
**Power:** Distributed, scalable high performance advanced analytics  
**Productivity:** Easier to build and deploy analytic applications  
**Enterprise Readiness:** Multi-platform

Our Services Deliver
**Knowledge:** Our experts enable you to be experts  
**Time-to-Value:** Our QuickStart projects give you a jumpstart  
**Guidance:** Our customer support team is here to help you

Our Philosophy
Customer-centric innovation  
Easy to do business with

Customers
200+ Global 2000

Global Presence
North America / EMEA / APAC

Global Industries Served
Financial Services  
Digital Media  
Government  
Health & Life Sciences  
High Tech  
Manufacturing  
Retail  
Telco
Solution: Revolution R Enterprise

Revolution R Enterprise is the **only** commercial big data analytics platform based on open source R statistical computing language.

- High Performance Analytics
- Big Data Analytics
- Cross-Platform
- Easier Build & Deploy
- Enterprise-Ready
Serving 200+ Data-Driven Enterprises

Finance & Insurance

Healthcare & Life Sciences

Academic & Gov’t

Consumer & Info Svcs

Manuf & Tech
## Comprehensive Partner Ecosystem

### Deployment / Consumption

- Oracle
- Business Objects
- KNIME
- Information Builders
- Rackspace
- Nebula
- QlikView
- alteryx
- Tableau
- Cognos
- MicroStrategy
- Amazon.com

### Advanced Analytics

- Revolution Analytics

### ETL

- alteryx
- Talend
- Informatica

### Data / Infrastructure

- Teradata
- Netezza
- IBM
- Greenplum
- AllianceONE
- Oracle
- AllianceONE
- IBM
- Microsoft
- Sybase
- Red Hat
- Vertica
- Honeywell
- Paraccel
- SAP
- Amazon.com

---

**SI / Services**

- Accenture
- Wipro
- Symphony Telesca
- Think Big Analytics
- Intel
- Big Computing
- IBM
- InfoGate
- Capgemini
- Deloitte
- Corio
- UO
- NEC
- OGIS-RI
- Infosys
- Cognizant

**Marketing Service Providers**

- ACXIOM
- Merkle
- Webb Will Mason
- Sapient

**Data Service Providers**

- Experian
- LexisNexis
QUICK INTRO TO R AND HADOOP
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

Download the White Paper
R is Hot
bit.ly/r-is-hot
R is open source and drives analytic innovation but has some limitations for Enterprises

- Memory Bound
- Single Threaded
- Community Support
- Innovative – 5000 packages+, exponential growth

<table>
<thead>
<tr>
<th>R</th>
<th>Revolution Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data</td>
<td>Bigger data sizes</td>
</tr>
<tr>
<td>Scale out, parallel processing, high speed</td>
<td>Speed of analysis</td>
</tr>
<tr>
<td>Commercial production support</td>
<td>Production support</td>
</tr>
<tr>
<td>Combines with open source R packages where needed</td>
<td>Innovation and scale</td>
</tr>
</tbody>
</table>
Revolution R Enterprise
High Performance, Multi-Platform Analytics Platform

Revolution R Enterprise

DeployR
Web Services
Software Development Kit

DevelopR
Integrated Development Environment

ConnectR
High Speed & Direct Connectors
Teradata, Hadoop (HDFS, Hbase), SAS, SPSS, CSV, OBDC

ScaleR
High Performance Big Data Analytics
Platform LSF, MS HPC Server, MS Azure Burst, SMP Servers

DistributedR
Distributed Computing Framework
Platform LSF, MS HPC Server, MS Azure Burst

RevoR
Performance Enhanced Open Source R+ CRAN packages
IBM PureData (Netezza), Platform LSF, MS HPC Server, MS Azure Burst, Cloudera, Hortonworks, IBM Big Insights, Intel Hadoop, SMP servers

Revolution Analytics Value-Add Components
Providing Power and Scale to Open Source R

Open Source R
Plus
Revolution Analytics performance enhancements
Open Source R performance: Multi-threaded Math

Customers report 5-50x performance improvements compared to Open Source R — without changing any code

<table>
<thead>
<tr>
<th>Computation (4-core laptop)</th>
<th>Open Source R</th>
<th>Revolution R</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear Algebra</strong>¹</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix Multiply</td>
<td>176 sec</td>
<td>9.3 sec</td>
<td>18x</td>
</tr>
<tr>
<td>Cholesky Factorization</td>
<td>25.5 sec</td>
<td>1.3 sec</td>
<td>19x</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>189 sec</td>
<td>74 sec</td>
<td>3x</td>
</tr>
<tr>
<td><strong>General R Benchmarks</strong>²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Benchmarks (Matrix Functions)</td>
<td>22 sec</td>
<td>3.5 sec</td>
<td>5x</td>
</tr>
<tr>
<td>R Benchmarks (Program Control)</td>
<td>5.6 sec</td>
<td>5.4 sec</td>
<td>Not appreciable</td>
</tr>
</tbody>
</table>

RevoScaleR brings the power of Big Data to R

Parallel External Memory Algorithms exploit available compute resources (cores & computers) independent of platform

Use the high-speed local data mart (XDF), or stream data from SAS, ODBC, HDFS or other remote data sources.

Abstracted communications layer provides portability of code between platforms: server, cluster, or in-database

Familiar, high-productivity programming environment for R users
A Revolution R Enterprise ScaleR analytic is provided a data source as input.

The analytic loops over data, reading a block at a time.

Blocks of data are read by a separate worker thread (Thread 0).

Worker threads (Threads 1..n) process the data block from the previous iteration of the data loop and update intermediate results objects in memory.

When all of the data is processed a master results object is created from the intermediate results objects.
Revolution R Enterprise ScaleR: High Performance Big Data Analytics

Data Prep, Distillation & Descriptive Analytics

R Data Step
- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- 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:
High Performance Big Data Analytics

Statistical Modeling

- **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 & Correlation Matrices**
- **Logistic Regression**
- **Classification & Regression Trees**
- **Predictions/scoring for models**
- **Residuals for all models**

Machine Learning

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

- **Variable Selection**
  - Stepwise Regression (for linear reg)

- **Simulation**
  - Monte Carlo

- **Cluster Analysis**
  - **K-Means**

- **Classification**
  - Decision Trees
ScaleR Performance and Capacity

GLM 'Gamma' Simulation Timings
Independent Variables: 2 factors (100 and 20 levels) and one continuous

Open Source R: glm()
Single threaded and RAM-intensive

RevoScaleR: rxGlm()
Fast, parallelized, and scalable

Timings from a Windows 7, 64-bit quadcore laptop with 8 GB RAM
Apache Hadoop

**Apache Hadoop** is an open source platform for data storage and processing that is…

✓ Scalable
✓ Fault tolerant
✓ Distributed

**Provides storage and computation in a single, scalable system.**
MapReduce

• MapReduce distributes jobs across nodes of the Hadoop cluster
• The Map function operates on a block of data and produces intermediate output
• The Reduce function takes the intermediate output and aggregates it into a final set of results.
• MapReduce jobs can be written in R, Pig, Java, Python and other languages
## Core Values of Hadoop

**A platform for all your data.**

<table>
<thead>
<tr>
<th>1</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Designed</strong> to manage data at petabyte scale</td>
</tr>
<tr>
<td></td>
<td><strong>Scale-out architecture</strong> increases capacity and processing power linearly</td>
</tr>
<tr>
<td></td>
<td><strong>Perform operations in parallel across the entire cluster</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Store data in any format</strong> – free from rigid schemas</td>
</tr>
<tr>
<td></td>
<td><strong>Define context at the time you ask the question</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Process and analyze data using virtually any programming language</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Build out your cluster on your hardware of choice</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Open source software guards against vendor lock-in</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Wide integration ensures investment protection</strong></td>
</tr>
</tbody>
</table>
REVOLUTION R ENTERPRISE
WITH

R AND HADOOP TOGETHER: OPTIONS
Why R and Hadoop?

- **Hadoop** offers a scalable infrastructure for processing massive amounts of data
  - Storage – HDFS, HBASE
  - Distributed Computing - MapReduce
- **R** is a statistical programming language for developing advanced analytic applications
- There is a need for more than counts and averages on these big data sets
- **Analyzing** all of the data can lead to insights that sampling or subsets can’t reveal.
Motivation for this project (RHadoop)

- **Make** it easy for the R programmer to interact with the Hadoop data stores and write MapReduce programs
- **Ability** to run R on a massively distributed system without having to understand the underlying infrastructure
- **Keep** statisticians focused on the analysis and not the implementation details
- **Open** source to drive innovation and collaboration.
A Common R-Based Analytic Platform Across Big Data Architectures

- Hadoop
- Linux or MSFT File-Based Cluster
- Big data appliances

Revolution Confidential
REVOLUTION ANALYTICS’ HADOOP CAPABILITIES: TODAY
Revolution R Enterprise server is physically connected to Hadoop cluster and is used to:

- Kick off MapReduce jobs using rmr
- Access Hadoop jar files (i.e. can submit a job to the job tracker)
- Connect to Hadoop data stores (HDFS and HBASE) using rhdfs or rhbase connectors
- Build and test models
- Collect results for further processing, visualization
- Propagate results through DeployR
Revolution R Enterprise

INSIDE Architecture

Use storage and processing power of data layer for:

- Data distillation
- Real-time R-based scoring
- Custom small-data parallel programming
- Late 2013: Big-data predictive models with ScaleR
Revolution Analytics’ Hadoop Capabilities

- Revolution R Enterprise 6.2 uses HDFS as a data source for RevoScaleR capabilities
- Hadoop Project
  - Revolution Analytics conceived, engineered and built the packages in the RHadoop Project
  - New releases available approximately every quarter
  - RHadoop Project consists of 3 R packages
    - `rhdfs` – connector from R to HDFS (read/write)
    - `rhbase` – connector from R to HBASE (read/write)
    - `rmr` – execute MapReduce jobs written 100% in R
ConnectR for Hadoop: Stream data from Hadoop to Revolution R Enterprise
Rhadoop: Map-Reduce with R

- Unlock data in Hadoop using only the R language
- No need to learn Java, Pig, Python or any other language
Manipulate HDFS directly from R
- Mimic as much of the HDFS Java API as possible

Examples:
- Read a HDFS text file into a data frame.
- Serialize/Deserialize a model to HDFS
- Write an HDFS file to local storage
  - rhdfs/pkg/inst/unitTests
  - rhdfs/pkg/inst/examples
rhdfs Functions

- **File Manipulations** - `hdfs.copy`, `hdfs.move`, `hdfs.rename`, `hdfs.delete`, `hdfs.rm`, `hdfs.del`, `hdfs.chown`, `hdfs.put`, `hdfs.get`
- **File Read/Write** - `hdfs.file`, `hdfs.write`, `hdfs.close`, `hdfs.flush`, `hdfs.read`, `hdfs.seek`, `hdfs.tell`, `hdfs.line.reader`, `hdfs.read.text.file`
- **Directory** - `hdfs.dircreate`, `hdfs.mkdir`
- **Utility** - `hdfs.ls`, `hdfs.list.files`, `hdfs.file.info`, `hdfs.exists`
- **Initialization** – `hdfs.init`, `hdfs.defaults`
rhbase

- Manipulate HBASE tables and their content
- Uses Thrift C++ API as the mechanism to communicate to HBASE

Examples
- Create a data frame from a collection of rows and columns in an HBASE table
- Update an HBASE table with values from a data frame
  - rhbase/pkg/inst/unitTests
rhbase Functions

- Table Manipulation – hb.new.table, hb.delete.table, hb.describe.table, hb.set.table.mode, hb.regions.table
- Row Read/Write - hb.insert, hb.get, hb.delete, hb.insert.data.frame, hb.get.data.frame, hb.scan
- Utility - hb.list.tables
- Initialization - hb.defaults, hb.init
Designed to be the simplest and most elegant way to write MapReduce programs

Gives the R programmer the tools necessary to perform data analysis in a way that is “R” like

Provides an abstraction layer to hide the implementation details

Examples

- Simulations - Monte Carlo and other Stochastic analysis
- R ‘apply’ family of operations (tapply, lapply…)
- Binning, quantiles, summaries, crosstabs and inputs to visualization (ggplot, lattice).
- Machine Learning
- rmr/pkg/inst/tests
rnr mapreduce Function

- `mapreduce (input, output, map, reduce, ...)`

  - `input` – input folder
  - `output` – output folder
  - `map` – R function used as map
  - `reduce` – R function used as reduce

  ... - other advanced parameters
Two Basic Deployment Models

- **Option 1** – rmr in use. **Revolution R Enterprise next-to and Revolution R Enterprise** installed *Inside* Hadoop to provide both:
  - rmr-enabled statistical analytics within Hadoop
  - rmr-enabled data distillation within Hadoop for statistical analyses inside or next to Hadoop

- **Option 2** – rmr not in use. **Revolution R Enterprise installed next-to** Hadoop to provide:
  - rhdfs- and rhbase-based access to Hadoop as a data source for HPA
  - statistical analysis done in Revolution R Enterprise on one or more edge nodes.
Option 1 – rmr in use

- Data distillation or Statistical Analysis are run as a MapReduce job in the Hadoop Cluster via rmr.
- Standalone Revolution R Enterprise “client” server (or cluster) is physically connected to Hadoop cluster and is used to:
  - Kick off MapReduce jobs using rmr
  - Access Hadoop jar files (i.e. can submit a job to the job tracker)
  - Connect to Hadoop data stores (HDFS and HBASE) using rhdfs or rhbase packages
  - Build and test models
  - Collect results for further processing, visualization
  - Propagate results through RevoDeployR.
- RRH  Is Revolution R Enterprise for Hadoop
When Working with Hadoop, Both Steps of Data Analysis Can Use MapReduce with rmr

- **Data Distillation/ Data Step**
  - rmr can be used within Hadoop to extract meaning from unstructured data
    - Create new variables such as counts (e.g. number of clicks in a day)
    - sort (e.g. according to criteria or sentiment)
    - merge
  - These sorts, merges, new variables, etc. can either be used within Hadoop for analytics or can be pulled into Revolution R Enterprise for statistical analysis

- **Statistical Analysis within Hadoop**
  - HPC-type analytics can be executed using rmr and R functions
  - HPA-type analytics can be executed using rmr via custom R scripting.
    - A library of RevoScaleR HPA routines for Hadoop is coming
Inside Hadoop: Data Distillation with R

- Log Files
- Event Streams
- Scrapped Text

Unstructured Data

HDFS Load

Structured Data

Map-Reduce

Records
Option 1 - Why have Revolution R on the Hadoop nodes?

- Improved Performance
  - MKL greatly accelerates matrix and vector math
- Enterprise Support
- Roadmap to RevoScaleR for Hadoop
- Simplified installation
- Roadmap:
  - Improved documentation
  - RHadoop available through Revolution Analytics
  - New, enterprise features in RHadoop
Option 2 – no rmr in use

- Hadoop data accessed from RRE using rhbase, rhdfs, RODBC (HIVE). We assume that rmr has not been used to distill / prepare the data
- Statistical analytics processing is on separate server or shared cluster using Revolution R Enterprise
Revolution with Hadoop Value Proposition

- Revolution Analytics is leading the movement to run R within Hadoop environments
- Revolution Analytics software has tremendous value as a complement to Hadoop for data distillation, statistical analysis, or both
- Install Revolution Analytics inside Hadoop for R-based Map Reduce and maximum data access flexibility
- Install Revolution R Enterprise on edge nodes to leverage Hadoop as a data source for HPA or HPC
- Quick Start programs to ensure a successful POC or first project are available now
Final thoughts

- R and Hadoop together offer innovation and flexibility needed to meet analytics challenges of big data
- Connects the R Programmer and the Hadoop Expert
- We need contributors to this project!
  - Developers
  - Documentation
  - Use cases
  - General Feedback
Resources

- Open source project: [https://github.com/RevolutionAnalytics/RHadoop/wiki](https://github.com/RevolutionAnalytics/RHadoop/wiki)
- Email: rhadoop@revolutionanalytics.com
REVOLUTION ANALYTICS’ HADOOP CAPABILITIES: SOON 😊
Hadoop in RRE v7

- RRE running “Inside” Hadoop
- Support for Cloudera and Hortonworks
  - No data movement
  - ScaleR inside Hadoop
  - WODA; just change Compute Context
- Other features:
  - Run inside Teradata 14.10; OSR 3; Decision Forests; PMML export; enhanced stepwise; others
CASE STUDIES
4X performance
50M+ records scored daily

**Key Technology:** Revolution R Enterprise and Hadoop, replacing SAS and Open Source R

**Outcomes:** Massively scalable infrastructure to support attribution and optimization at an individual customer level (segments of one) for clients such as Williams-Sonoma. Client saved $250K in one campaign.

Rapid development and deployment of customer-specific models, using innovative analytic techniques such as big data GAM Survival models

**Bottom Line:** Driving revenue lift and cost savings through marketing optimization

---

**Profile:** Multi-channel marketing attribution and analytics software developer and service provider. Growing, innovative, cost-conscious.

“Given that our data sets are already in the terabytes and are growing rapidly, we depend on Revolution R Enterprise’s scalability and power. We saw about a 4x performance improvement on 50 million records. It works brilliantly.”

CEO, John Wallace
(DataSong formerly named UpStream)
DataSong’s Revolution Analytics + Hadoop Integration

Current State: Revo v6

- Functions to read Hadoop output; xdf creation
- Exploratory data analysis
- GAM survival models

Upstream Data Format (UDF)

Data Inputs
- ETL
- N marketing channels
- Behavioral variables
- Promotional data
- Overlay data

Custom Variables (PMML)

Scoring
- Scoring for inference
- Scoring for prediction
- 5 billion scores per day per customer
High-Performance Analytics Consulting

24 hours to under 4 hours
Use more data
Iterate 100’s of models

Profile: More than 3,500 decision-science professionals helping companies institutionalize data-driven decision making and harness Big Data

“We use Revolution Analytics’ rmr package for R-powered big data analytics in Hadoop environments so that we can do production scoring at a granular level. It allows us to parallelize high speed computations on massive amounts of data.”

VP / Head of Innovation & Development, Zubin Dowlaty

Key Technology: Revolution R Enterprise and Rhadoop rmr (R MapReduce) package, replacing SAS & EDW

Outcomes: Built and deployed second-generation, high-performance, scalable analytics applications for Fortune 500 companies

Bottom Line: Ability to achieve first-mover advantage for clients in areas such as fraud analytics, loyalty program analytics, multi-channel marketing analytics and inventory/supply chain analytics
Recommendation engine generates high-resolution planting prescription to optimize growers’ yield

Profile: Data Scientists and IT at global agricultural products developer invent a novel, analytics-powered service that bolsters sales of other products and drives customer success

“We want to provide more than just prescription for planting - we want to make recommendation for products all year long. When we’re doing descriptive analytics, there’s no better tool than Revolution R Enterprise and rmr. Once you start using R, you find that the openness and flexibility are huge. We’ll be testing the strength of our models with historical data and continuously improving them, so we need very good performance on huge data sets.”

Key Technology: Revolution R Enterprise using ScaleR Big Data analytics capabilities and rmr (R MapReduce) in Cloudera’s CDH Hadoop distribution

Outcomes— Successful product launch powered by custom code combined with 15+ CRAN packages on 200 TB of data sourced from company’s seed yield research experiments, government and other 3rd-party data, and growers’ own data regarding yield history, land, conditions, etc.

Bottom Line: Revenue growth from new and existing products

Data Scientist in IT/ R&D
Revolution R Enterprise + Hadoop – Sample Solutions

- Fraud Detection
- Web click analysis
- Personalized offers
- Preventative maintenance
- Sentiment analysis
- Risk analysis
- What would you like?
Innovates to Outperform

**Profile:** Publicly-traded, investment management company adopts open, non-proprietary systems to “compete based on strength of analytics” for $8.5B portfolio of 22 funds.

“One of the first R-based production deployments we rolled out tracks revenue flows among manufacturers and their suppliers. We combine public and proprietary data and apply graph analyses to get a clearer understanding of the likely performance of suppliers. These forecasts are more accurate than what could be developed with quarters-old public financial reports.”

- Sr. Quantitative Researcher, Tal Sasani

**Key Technology:** Revolution R Enterprise replacing closed, proprietary industry applications. Tableau front end for production analytics.

**Outcomes:** Battery of custom strategy simulation & portfolio optimization now run overnight to inform morning work. RRE 20x faster than open source R.

**Put R-based analytics into production.**

**Bottom Line:** Company competes on strength of custom-built simulations, scenario analyses & financial stress tests.
Scales Business with Analytics

TB’s from 200+ data sources
10’s thousands attributes
100’s millions of scores daily

Key Technology: Revolution R Enterprise and
IBM PureData System for Analytics (powered by Netezza technology)

Outcomes: Moving R functions to the data in IBM reduces infrastructure costs by 50%
by avoiding need for redundant analytics architecture for model training & scoring

Profile: Smart data service provider helping marketers target online advertising to individuals
with the highest propensity to convert

“We’ve been able to scale our solution to a problem that’s so big that most companies
could not address it. If we had to go with a different solution we wouldn’t be as efficient as we are now.”

- SVP Analytics, Kevin Lyons

Client-custom, complex ensemble models built and scored daily on massive data

Bottom Line: Simplifying complicated big data analytics workflow enables massive scalability for in-demand marketing data
Next-Generation Predictive Optimization Engine Outperforms Legacy Predecessor

Profile: Software and services for optimized digital marketing through multi-channel visitor experiences on personalized websites and real-time digital audience targeting for clients such as Verizon, JP Morgan Chase, Intuit

“We need a high-performance analytics infrastructure because marketing optimization is a lot like a financial trading. By watching the market constantly for data or market condition updates, we can now identify opportunities for our clients that would otherwise be lost.”

- Chief Analytics Officer, Leon Zemel

>2x data
>2x attributes
Improved model performance

Key Technology: Revolution R Enterprise (replacing legacy analytics platform vendor), pulling data from EDW, informing rules engine in [x+1]’s proprietary production system

Outcomes: higher lift of real time multi-channel ad targeting analytics derived from use of more data and attributes -- new proprietary data (CRM systems, call center transcripts, company website) & 3rd party demographic & behavioral data

Bottom line: Drive higher lift through higher precision audience targeting and tailored messaging achieved through use of more data
Economic Capital Modeling

1 day to 15 minutes
100,000 years of simulations
Pricing optimization increases financial health

Key Technology: Revolution R Enterprise replaced Excel; drives business rules in company production system

Outcomes: Ability to compensate for lack of historical data by simulating a wide variety and quantity of events and using advanced correlation techniques. Complete full day of work in 15 minutes

Bottom line: Improved financial health by managing risk and increasing pricing optimization

Profile: 10-year-old reinsurer’s Actuarial Group systematically makes sound financial and pricing decisions in production system and completes ad hoc analysis.

“As things become more and more extreme, I need a model that can estimate my risk in a way to that enhances our confidence in our pricing and reserving. Modeling with Revolution R Enterprise gives me that.”

VP and Pricing Actuary, Jamie Botelho
Market Surveillance @ F100 Investment Co.

Profile: Full-service global investment and securities management firm’s IT team proved effectiveness of Revolution R Enterprise to detect potentially costly outliers and errors

Key Technology: Revolution R Enterprise using ScaleR Big Data Analytics capabilities

Analytic Approach – Exchange Rate Error Detection: ARIMA and VAR models used to define acceptable value changes using the prediction for the next value in a time series. Models trained using historical data.

Bottom line: new analytics paradigm for existing processes introduced, with potential for millions of dollars in cost avoidance

Analytic Approach – Outlier Detection: Use historical data for each customer (>65M end-of-day trades and >8,500 variables) to build and train linear regression model to establish range of predicted values for customers’ trades so that actual trades can be analyzed for outliers.

“Using statistical analysis by customer delivers superior accuracy compared to rules-based analysis (such as analyzing largest 10% of trades), which fail over time as volumes or client behavior changes. Statistical models that can be retrained (e.g. weekly) will account for changes and not fail over time.” VP, IT
Quantitative Research @ Global Investment Co.

Value at Risk (VAR) analysis deployed to 100’s of traders

Profile: Quantitative Research team does Value at Risk calculations to allow 100’s of traders to evaluate instruments on daily basis.

Key Technology: Revolution R Enterprise using ScaleR Big Data and DeployR integrated with Siteminder, which provides a secure, transparent, centralized analytics center.

Analytic Approach – develop models that can be applied to real-world data to exploit market opportunities and successfully develop, back-test, and deploy quantitative and event-based trading and investment strategies to generate alpha and effectively manage risk.

Challenge - Quantitative Research Group had “big data” and “big computation” challenges. Needed centralized, scalable, high-performance platform to measure value and risk of each instrument in the portfolio to inform daily trading decisions.

Bottom Line - Powerful statistical analytics platform provides centralized, secure model repository and guides hundreds of millions of dollars of transactions made by 100’s of traders.
Q & A