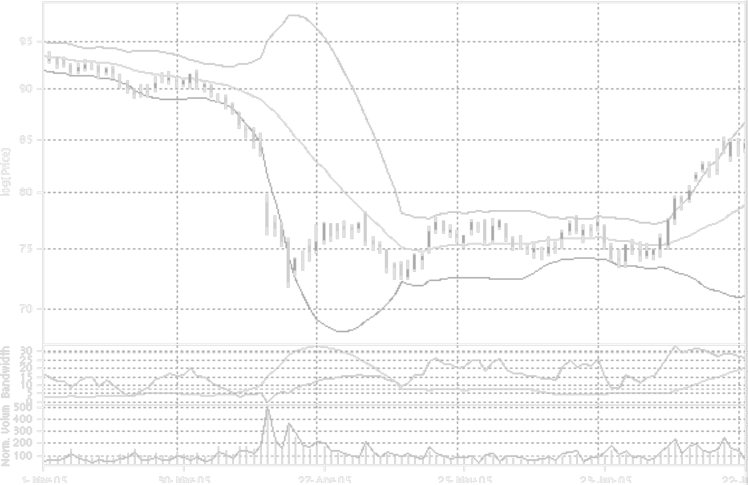
Revolution Analytics

Leveraging R in Hadoop Environments

September 21, 2011



In Today's Webinar:

- About Revolution Analytics
- Why R and Hadoop?
- The Packages (rhdfs, rhbase, rmr)
- Examples
- Resources and Further Reading

- Co-sponsored by Revolution and Cloudera

REVOLUTION ANALYTICS

OPEN SOURCE ANALYTICS FOR THE ENTERPRISE

- ➔Most advanced statistical analysis software available
- ➔Half the cost of commercial alternatives
- ➔2M+ Users
- ➔ 3,000+ Applications

Statistics

Predictive Analytics

Data Mining

Visualization

Finance

Life Sciences

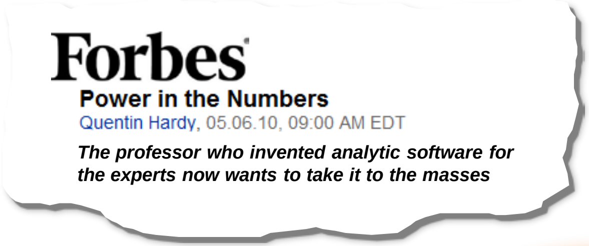
Manufacturing

Retail

Telecom

Social Media

Government

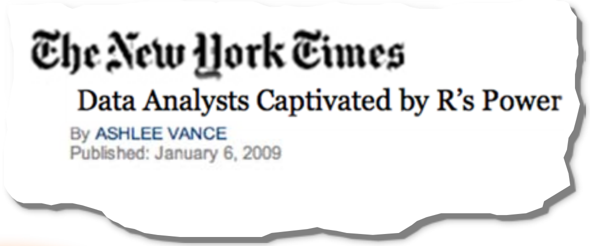


Forbes

Power in the Numbers

Quentin Hardy, 05.06.10, 09:00 AM EDT

The professor who invented analytic software for the experts now wants to take it to the masses

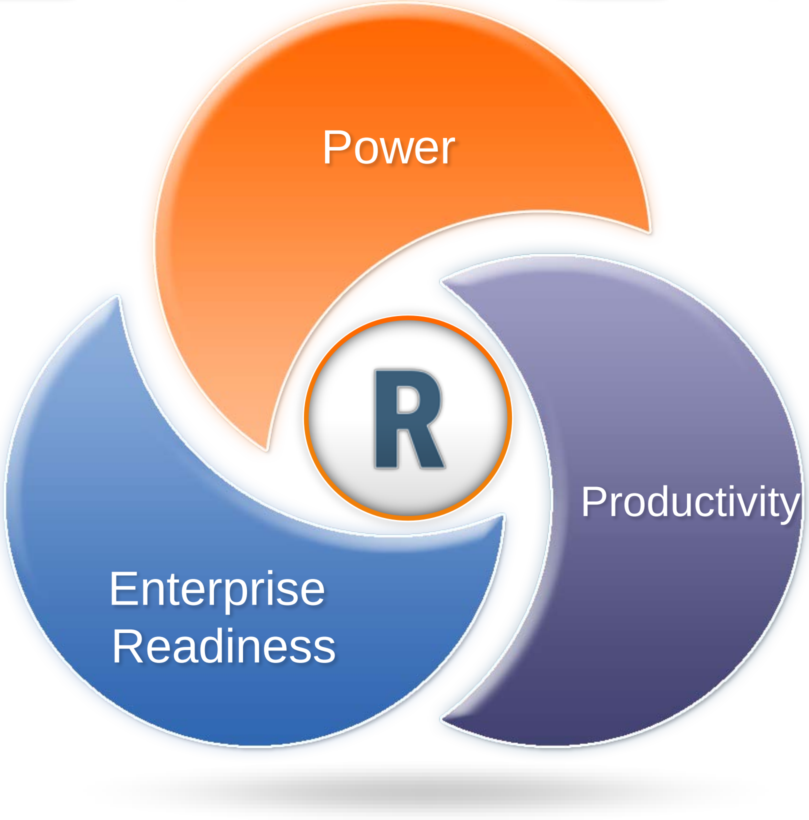


The New York Times

Data Analysts Captivated by R's Power

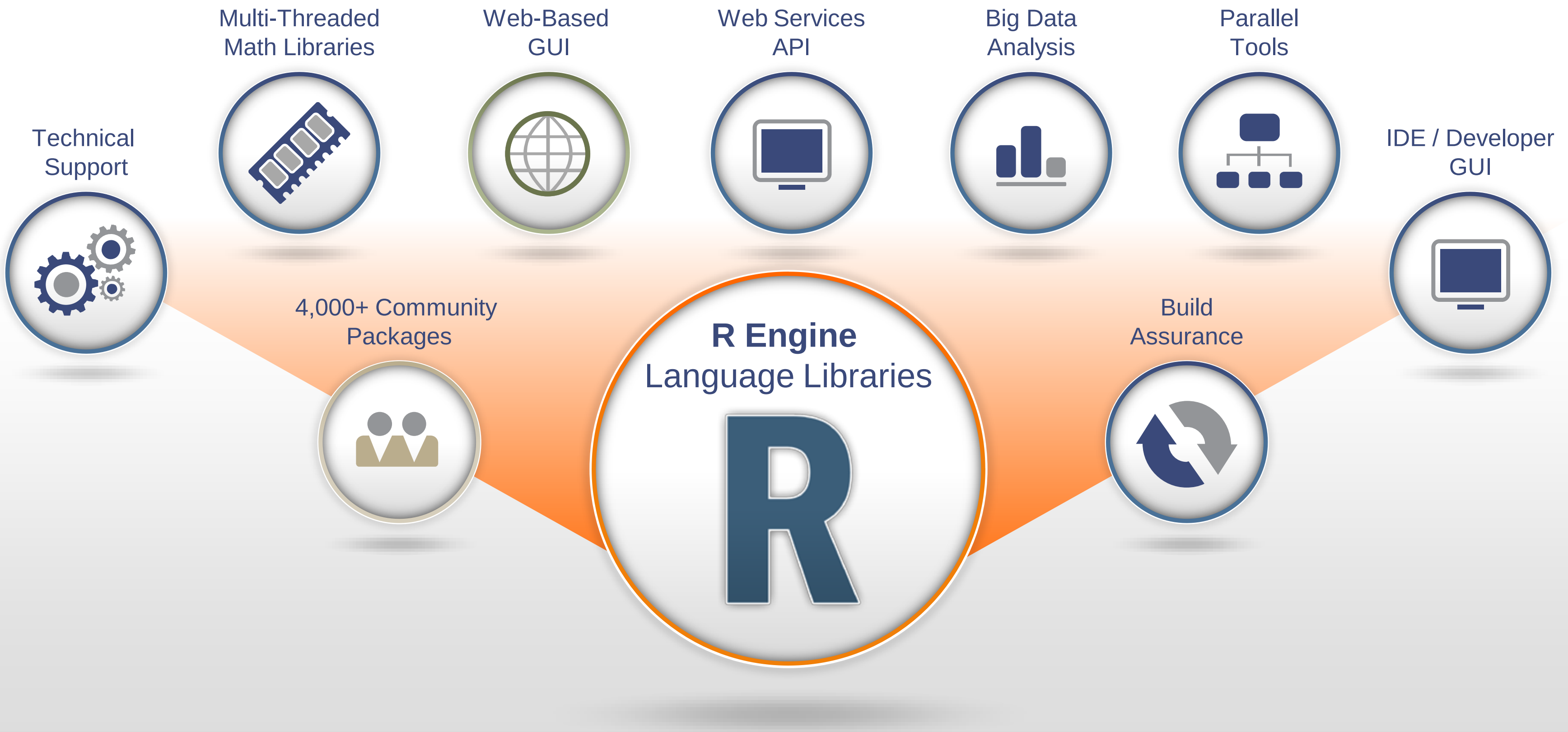
By ASHLEE VANCE

Published: January 6, 2009



What's the Difference Between R and Revolution R Enterprise?

Revolution R is 100% R and More®



Technical Support

Multi-Threaded Math Libraries

Web-Based GUI

Web Services API

Big Data Analysis

Parallel Tools

IDE / Developer GUI

4,000+ Community Packages

R Engine
Language Libraries

Build Assurance

For more information contact: info@revolutionanalytics.com

Let's Talk about R and Hadoop

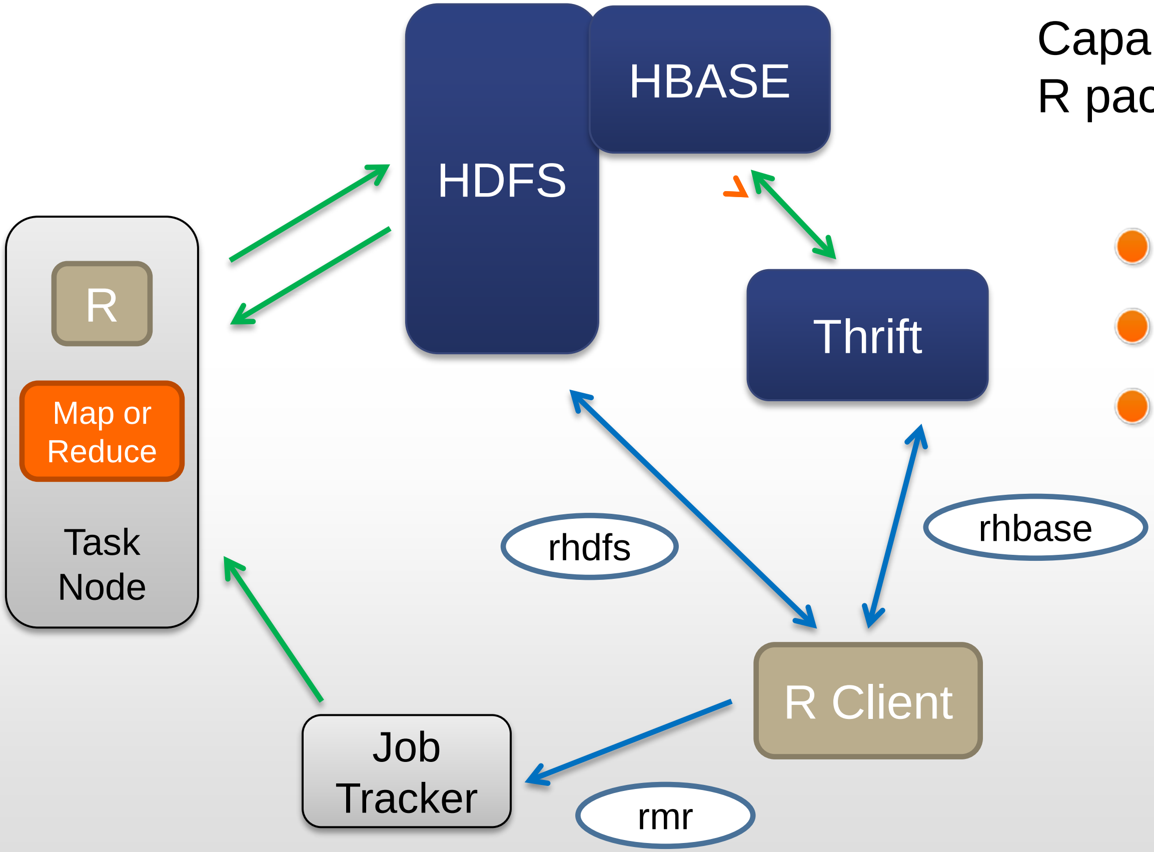
Why R and Hadoop?

- **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

- **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.

R and Hadoop – The R Packages



Capabilities delivered as individual R packages

- rhdfs - R and HDFS
- rhbase - R and HBASE
- rmr - R and MapReduce

Downloads available from Github

rhdfs

- 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

rmr

- 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).
  - Data Mining and Machine Learning
  - `rmr/pkg/inst/tests`

rmr 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

The Basics

```
small.ints = 1:10
out = lapply(small.ints, function(x) x^2)

small.ints = to.dfs(1:10)
out = mapreduce(input = small.ints,
                map = function(k,v) keyval(k, k^2))

groups = rbinom(32, n = 50, prob = 0.4)
out = tapply(groups, groups, length)

groups = to.dfs(groups)
out = mapreduce(input = groups,
                reduce = function(k,vv) keyval(k, length(vv)))
```

K-means

```
kmeans =
  function(points, ncenters, iterations = 10,
           distfun =
             function(a,b) norm(as.matrix(a-b), type='F')){
    newCenters = kmeans.iter(points, distfun = distfun, ncenters = ncenters)
    for(i in 1:iterations) {
      newCenters = lapply(values(newCenters), unlist)
      newCenters = kmeans.iter(points, distfun,
                               centers = newCenters)}
    newCenters}
```

```
kmeans.iter =
 function(points, distfun, ncenters = length(centers),
      centers = NULL) {
  from.dfs(
   mapreduce(input = points,
     map = if (is.null(centers)) {
          function(k,v)keyval(sample(1:ncenters,1),v)}
        else {
          function(k,v) {
           distances =  lapply(centers,   function(c)distfun(c,v))
          keyval(centers[[which.min(distances)]],v)}},

     reduce = function(k,vv)  keyval(NULL,apply(do.call(rbind,vv),2,mean))))}
```

Final thoughts

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

Resources

- Slides / Replay: bit.ly/r-and-hadoop
- Open source project: https://github.com/RevolutionAnalytics/RHadoop/wiki
- Participate in our survey: http://www.surveymonkey.com/s/JM3N6RP
- Revolution R Enterprise: bit.ly/Enterprise-R
- Cloudera CDH : http://www.cloudera.com/hadoop/
- Email: rhadoop@revolutionanalytics.com

Thank you.

Revolution Analytics

The leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com	650.330.0553	Twitter: @RevolutionR