

SQL Server 2016 and Microsoft R Server

Tomaž Kaštrun

March 16, 2017

SQL User Group



Speaker Info



- BI Developer and data analyst
- SQL Server, SAS, R, Python, C#, SAP, SPSS
- 15years experience MSSQL, DEV, BI, DM
- Spar ICS Austria, Spar Slovenija
- Frequent community speaker
- Avid coffee drinker & Bicycle junkie



<http://tomaztsql.wordpress.com>



tomaz.kastrun@gmail.com



@tomaz_tsq1



/in/tomaztsql



<http://github.com/tomaztk>



<https://mvp.microsoft.com/PublicProfile/5002196>





Agenda

- 1 R language and available RevoscaleR package for multi-threaded and parallelization computation
- 2 Using R language in T-SQL for data analysis and predictions
- 3 Visualizations (PowerBI)



Analytical Barriers

Common Challenges

Uncertain total cost of ownership	Inadequate access to important business data	Limited business agility	Limited business value

Addressing Challenges with R from Microsoft

Peace of mind	Efficiency	Speed and scalability	Flexibility and agility



What is R?

- A Language Platform
 - A Procedural Language optimized for Statistics and data science (and much more)
 - A Data Visualization framework
 - Provided as Free Software
- A Community and a system
 - Taught on universities and many active user groups across the world
 - Estimated 3Mio Users
 - Repositories (CRAN, BioConductor, Github,...)

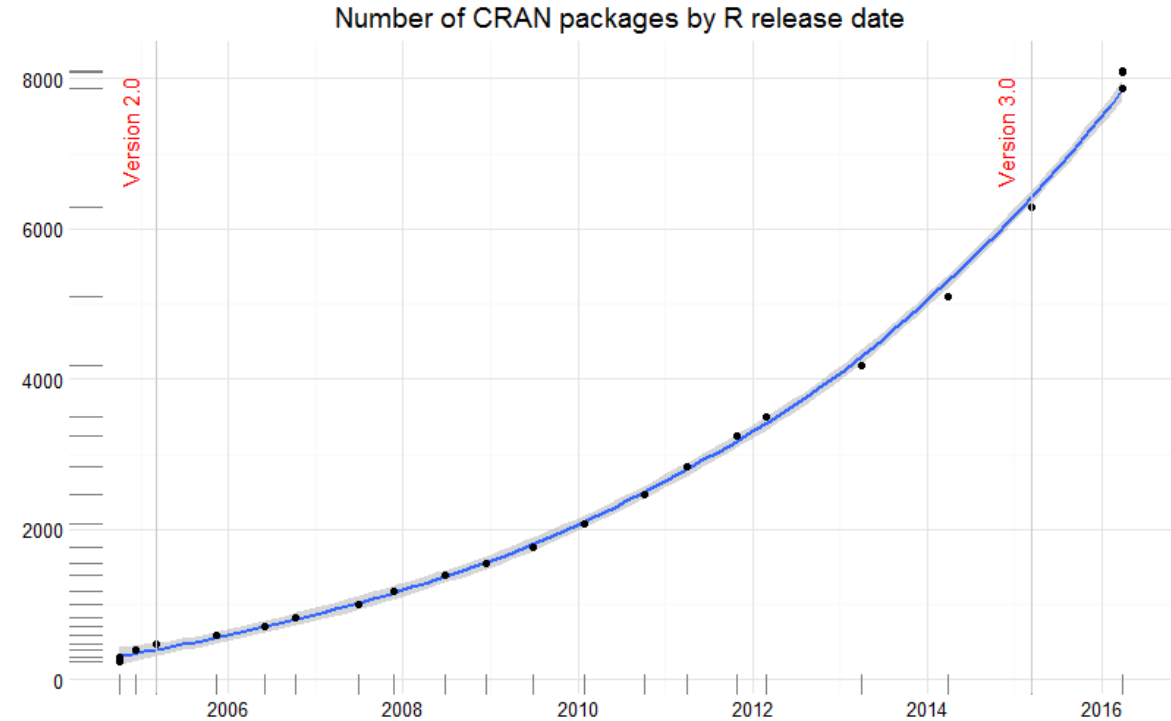
In fact, R is a movement!



Power of R: R Language + Packages



- R is an open source ([GNU](#)) version of the S language developed by John Chambers *et al.* at Bell Labs in 80's [History of R](#)
- R was initially written in early 1990's by [Robert Gentleman](#) and [Ross Ihaka](#) then with the Statistics Department of the University of Auckland
- R is administered and controlled by the [R Foundation](#)
- Microsoft is founding member and Platinum Sponsor of [R Consortium](#)



3000 packages added in last 2 years

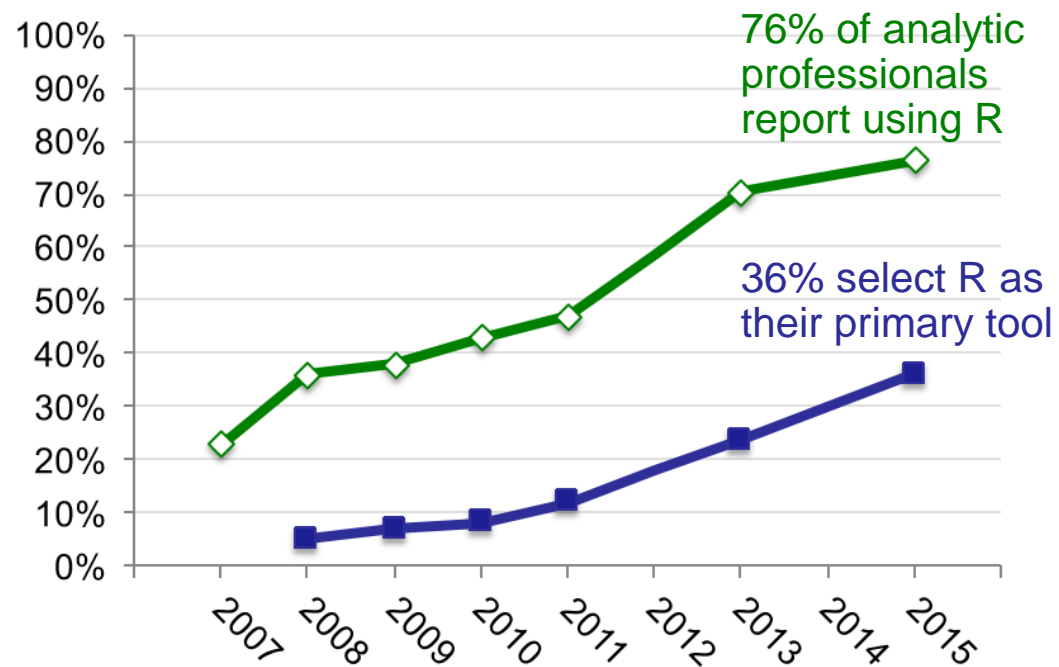


R: The #1 software for Advanced Analytics



R Usage Growth

[Rexer Data Miner Survey](#), 2007-2015



Language Popularity

[IEEE Spectrum Top Programming Languages](#), 2016

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

R is Popular

Most widely used software for Data Mining and Analytics

Used by 2M+ data scientists, statisticians and analysts

Open Source (GPL) language and environment

Easy to bring and explore data, uncover insights and generate predictions

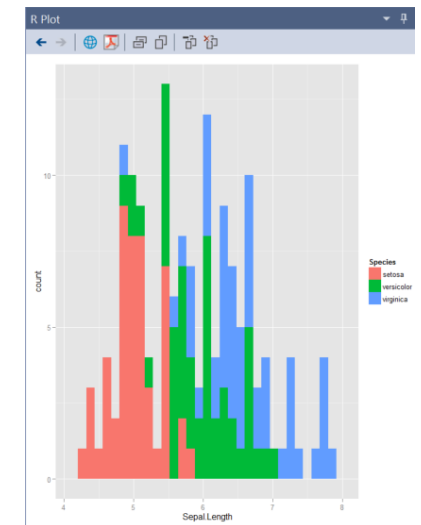
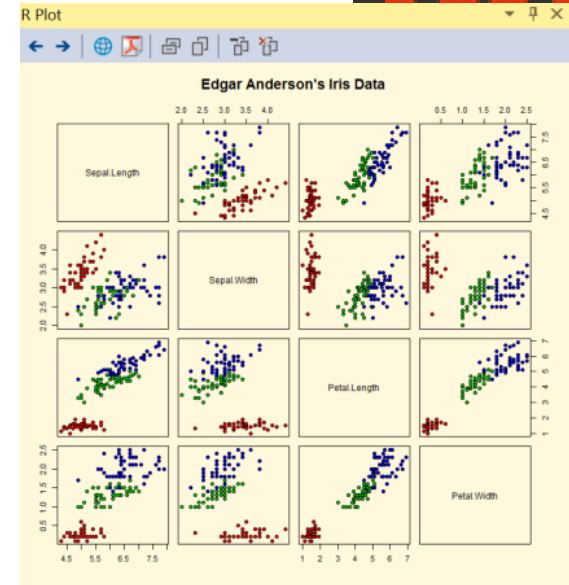
From the most trivial statistical function to the most complex ML technique

Easily create beautiful and unique data visualizations

As seen in New York Times, The Economist and FlowingData

Thriving open-source community

Over 8000 packages in CRAN and growing; Active forums and groups



Power of R: R Language + Packages

CRAN: 9000+ Add-on packages for R



CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know have no idea they exist. As an effort to make them more widely known I thought I'd jazz up the index page. Images are free to use, and got from [iStock](#) stock photo site. Visual puns are mine. Task View links go to the cran.r-project.org site and not a mirror.



Bayesian Inference

Applied researchers interested in Bayesian statistics are increasingly attracted to R because of the ease of which one can code algorithms to sample. [\[more\]](#)



Chemometrics and Computational Physics

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of. [\[more\]](#)



Clinical Trial Design, Monitoring, and Analysis

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including. [\[more\]](#)



Cluster Analysis & Finite Mixture Models

This CRAN Task View contains a list of packages that can be used for finding groups in data and modelling unobserved cross-sectional heterogeneity. Many... [\[more\]](#)



Probability Distributions

For most of the classical distributions, base R provides probability distribution functions (p), density functions (d), quantile functions (q), and. [\[more\]](#)



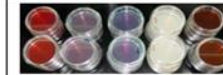
Computational Econometrics

Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... [\[more\]](#)



Analysis of Ecological and Environmental Data

This Task View contains information about using R to analyse ecological and environmental data. [\[more\]](#)



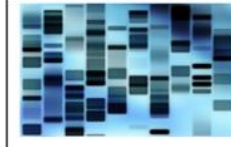
Design of Experiments (DoE) & Analysis of Experimental Data

This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements... [\[more\]](#)



Empirical Finance

This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic. [\[more\]](#)



Statistical Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs). [\[more\]](#)



Natural Language Processing

This CRAN task view contains a list of packages useful for natural language processing. [\[more\]](#)



Analysis of Pharmacokinetic Data

The primary goal of pharmacokinetic (PK) data analysis is to determine the relationship between the dosing regimen and the body's exposure to the drug as. [\[more\]](#)



Official Statistics & Survey Methodology

This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide... [\[more\]](#)



Phylogenetics, Especially Comparative Methods

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical. [\[more\]](#)



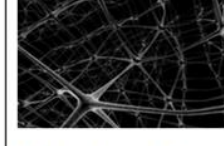
Multivariate Statistics

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this... [\[more\]](#)



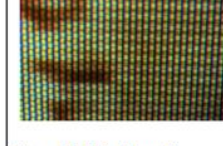
Optimization and Mathematical Programming

This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics. [\[more\]](#)



Machine Learning & Statistical Learning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually. [\[more\]](#)



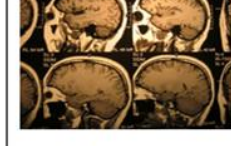
Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic. [\[more\]](#)



High-Performance and Parallel Computing with R

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are. [\[more\]](#)



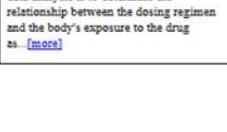
Medical Image Analysis

This task view is for input, output, and analysis of medical imaging files... [\[more\]](#)



Analysis of Spatial Data

Base R includes many functions that can be used for reading, visualizing, and analysing spatial data. The focus in this view is on "geographical" spatial. [\[more\]](#)



Survival Analysis

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an. [\[more\]](#)



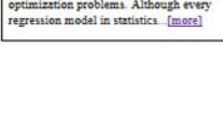
Time Series Analysis

Base R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are... [\[more\]](#)



Robust Statistical Methods

Robust (or "resistant") methods for statistics modelling have been available in S from the start, in R in package stats (e.g., median(), mean(), trim =). [\[more\]](#)



Statistics for the Social Sciences

Social scientists use a wide range of statistical methods. To make the burden carried by this task view lighter, I have suppressed detail in some areas that... [\[more\]](#)



gRaphical Models in R

Wikipedia defines a graphical model as a graph that represents dependencies among random variables by a graph in which each node is a random variable, and. [\[more\]](#)



Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better. [\[more\]](#)



Psychometric Models and Methods

Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked... [\[more\]](#)

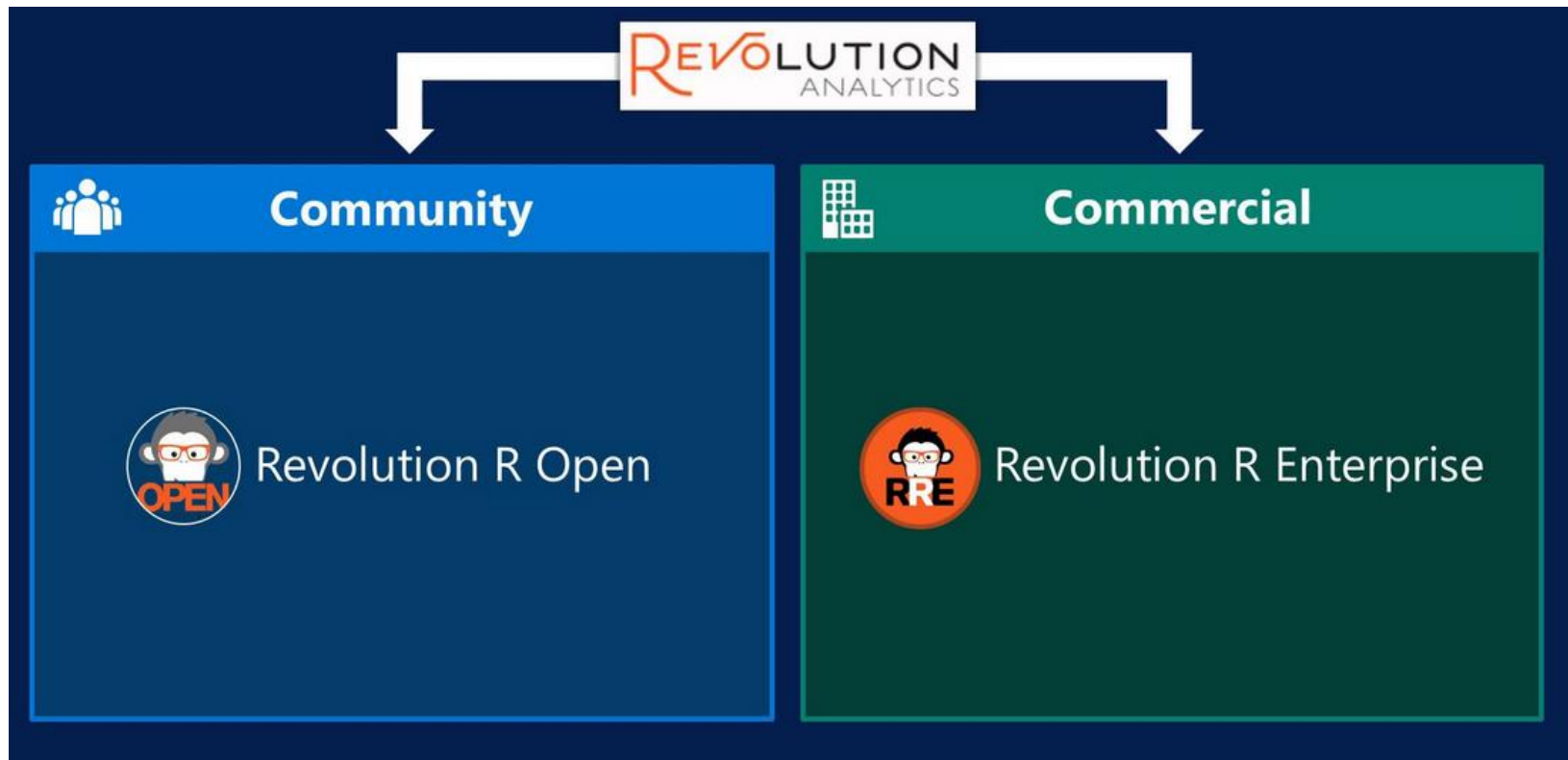


Limitations of R as a free software

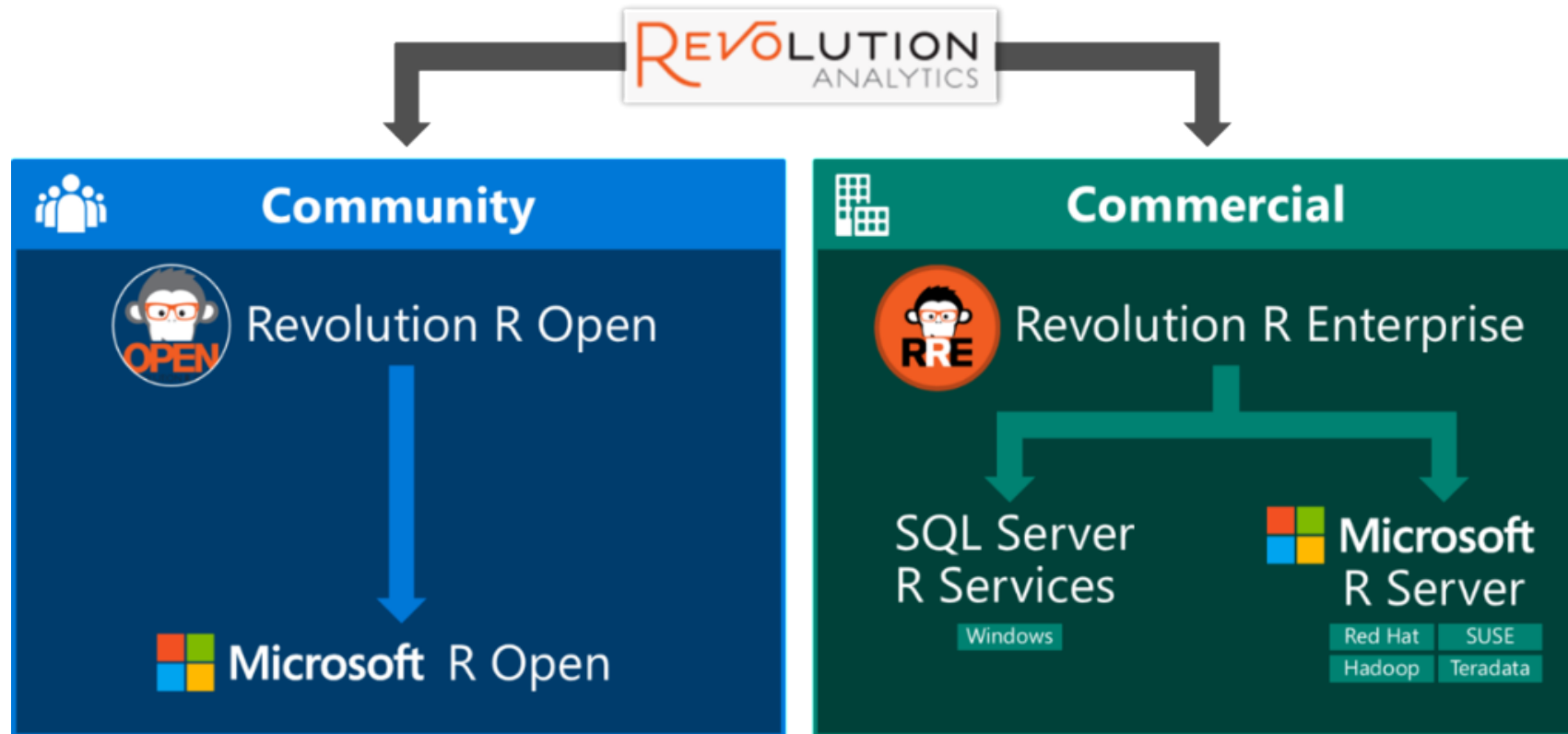
- Memory Based Data access model
- Interpreted vs. Compiled Performance
- Lack of parallel computation
- Data movement & Duplication Costs
- Governance and providence oversight
- Community support vs. Enterprise utilization



Revolution Analytics Product Integration



Microsoft R SQL Server platform



- > Free and open Source R distribution
- > Enhanced and distributed by Revolution analytics

- > Built in Advanced Analytics and Standalone Server Capability
- > Leverages the benefits of SQL Server 2016EE

Microsoft R Platform



Microsoft R Open

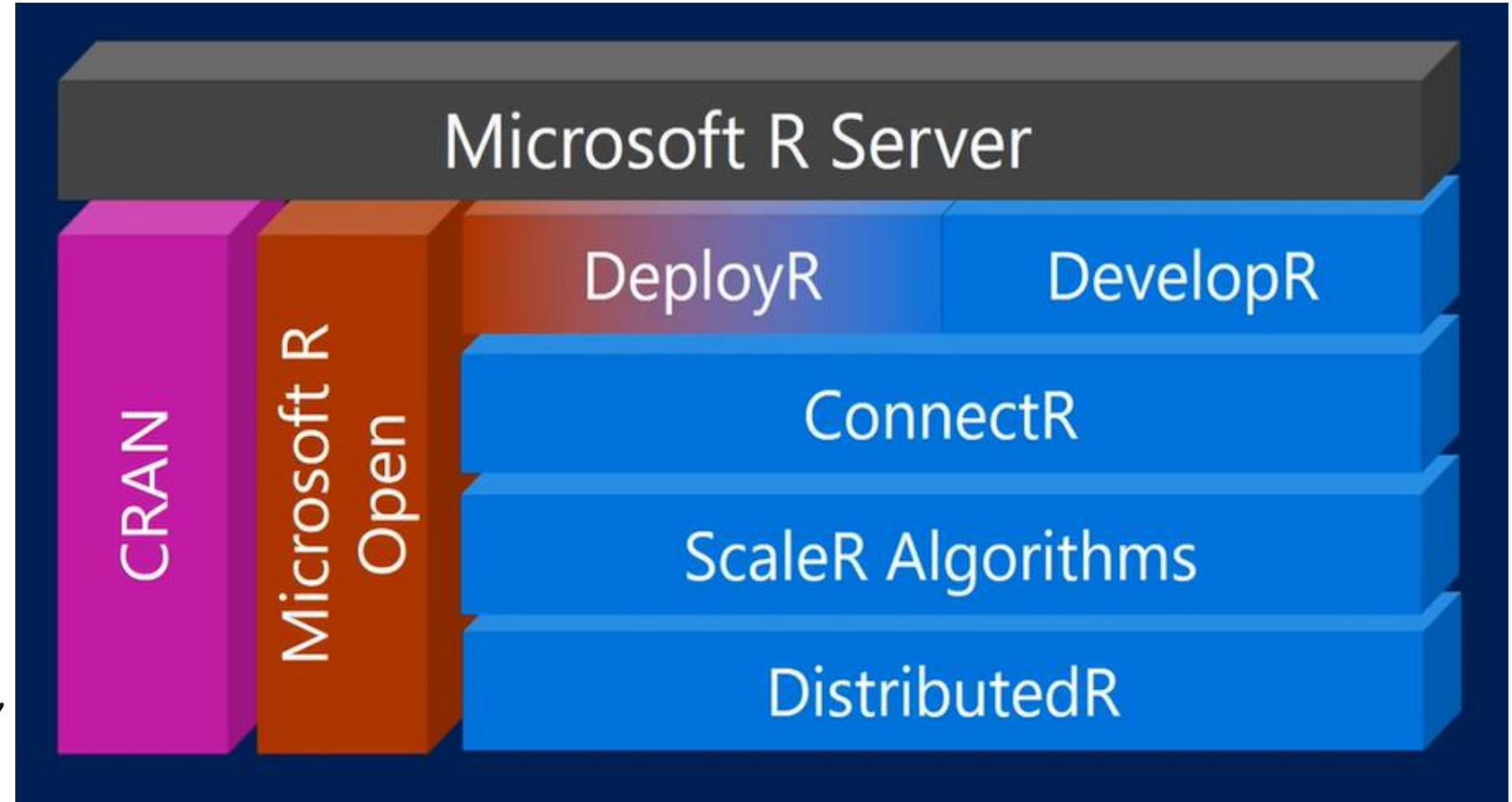
Microsoft R Client

Microsoft SQL R Services

Microsoft R Server

Different flavors:

Microsoft R server for Linux,
Microsoft R Server for Teradata,
Microsoft R Server for Hadoop,
Microsoft R HDInsight





Microsoft R Server

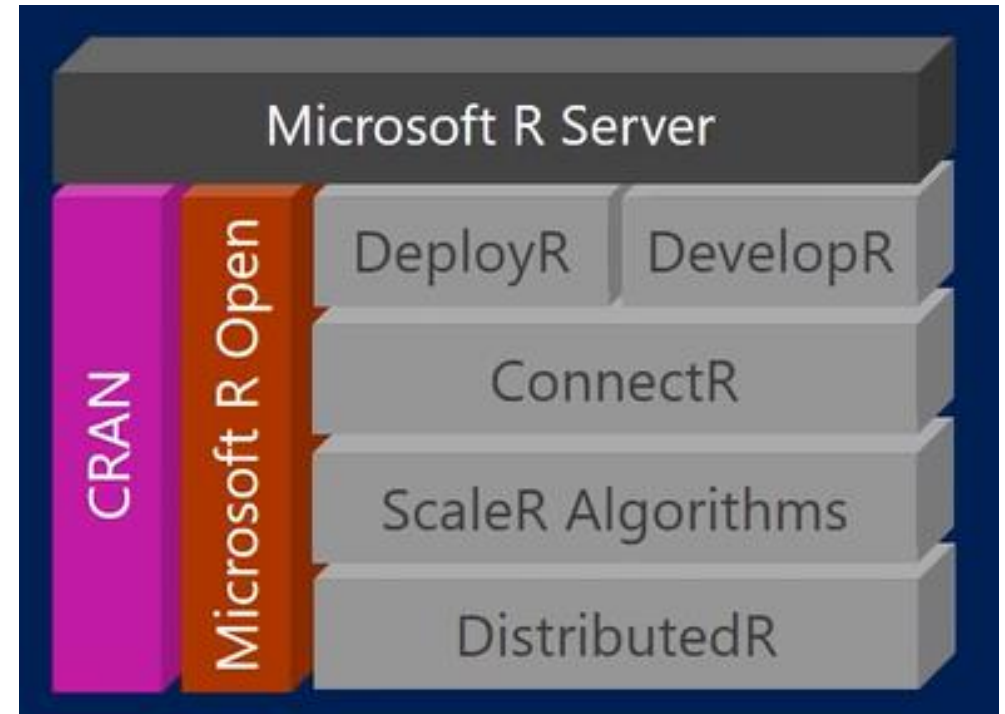
- Evolved from Revolution R Enterprise
- Based on open Source R
- Adapted for Enterprise Scale
- For multiple platforms
 - Hadoop
 - Teradata
 - Linux
 - Azure
 - Windows
- Interoperable
- On-premises + Cloud + Hybrid
- Operationalize analytics for Big scale datasets and big data





Based on Open Source R

- Open source based
- Runs your normal R Script
- MetaCran / CRAN / Github / Bioconductor

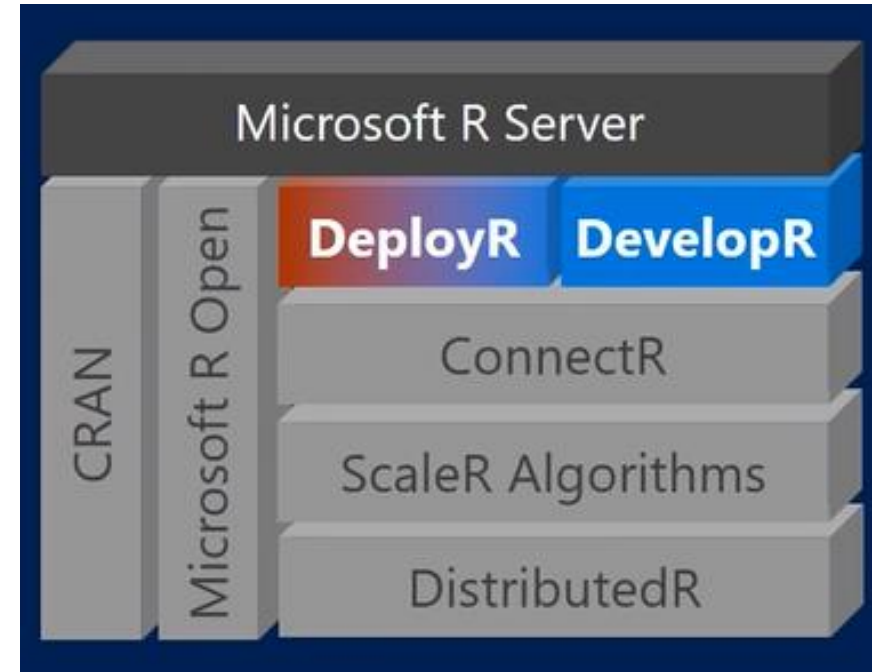


Microsoft R Server



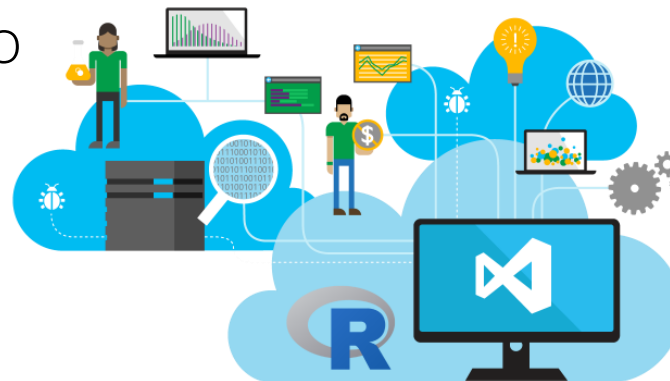
DeployR

- Web service - API integration
- Compatible with array of tools
- Abstract usage of R without knowing it



DevelopR

- R IDE based on Visual studio
- Rstudio for linux Users
- Client Based



Microsoft R Server



ConnectR

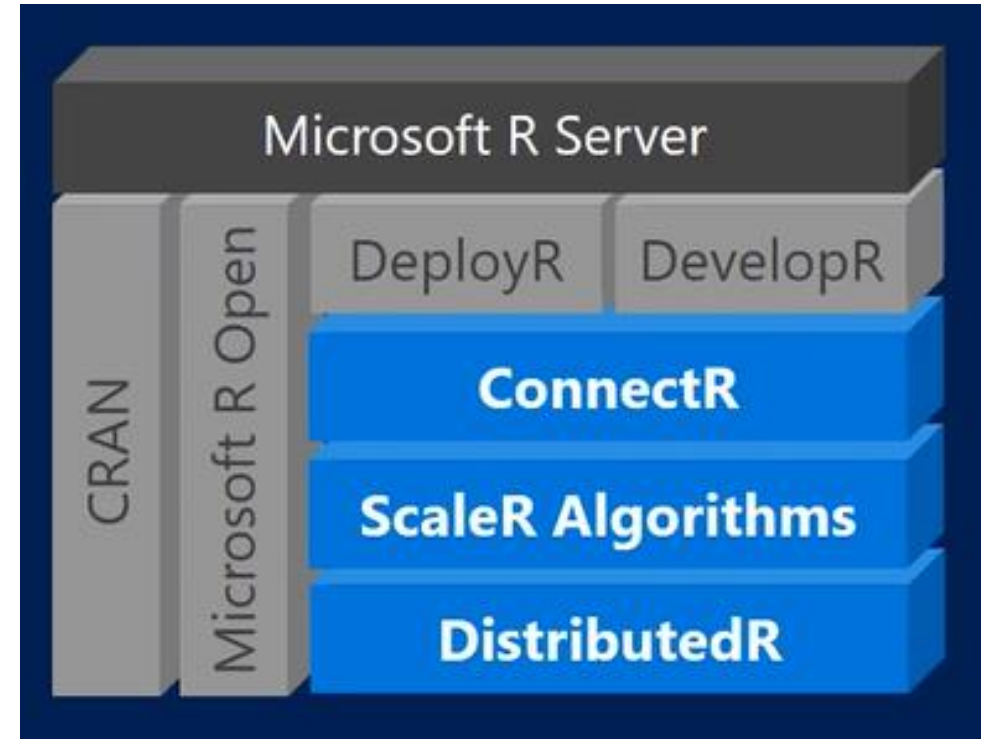
- Series of connectors for consistent access to scaleR algorithms

DistributedR

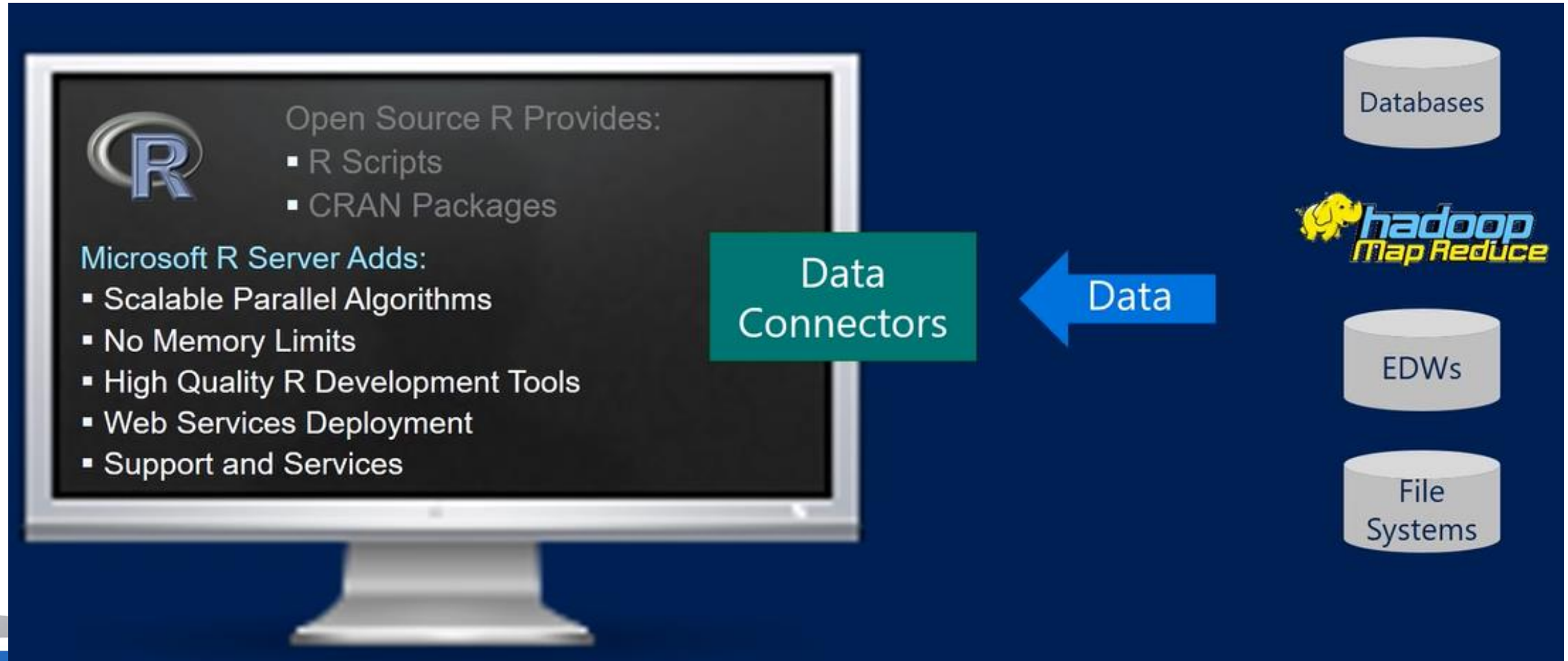
- Normalization layer for ScaleR algorithms (SQLServer, Win, Lin, TeraData, Hadoop, HDI)

ScaleR

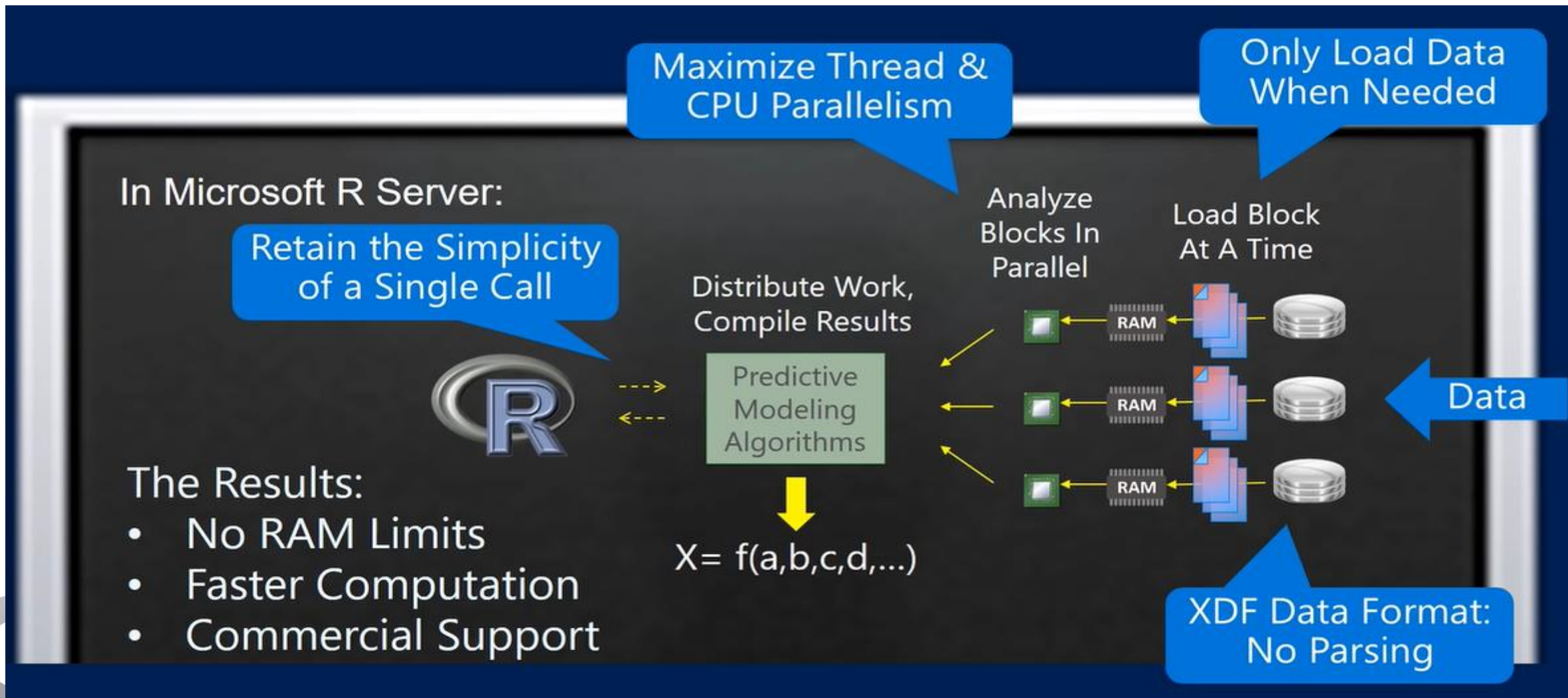
- Typical statistical approaches refactored for parallel computation
- Block-wise computation; No In-Memory constraints



Microsoft R Server - architecture



Parallelizing data process





How does parallelization work

In Microsoft R Server:



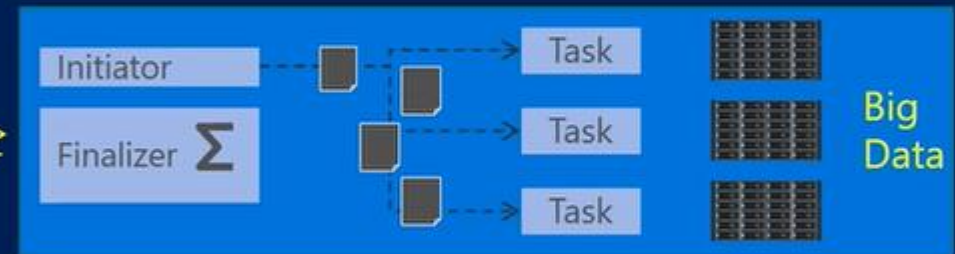
- ... load a large dataset into Hadoop
- ```
CContext <- RxInHadoopMR()
rxSetComputeContext(CContext)
Model_obj <- rxLinMod(...)
```
- ... use model object to predict...

Stub

Remote Algorithms

Moves Logic To The Data

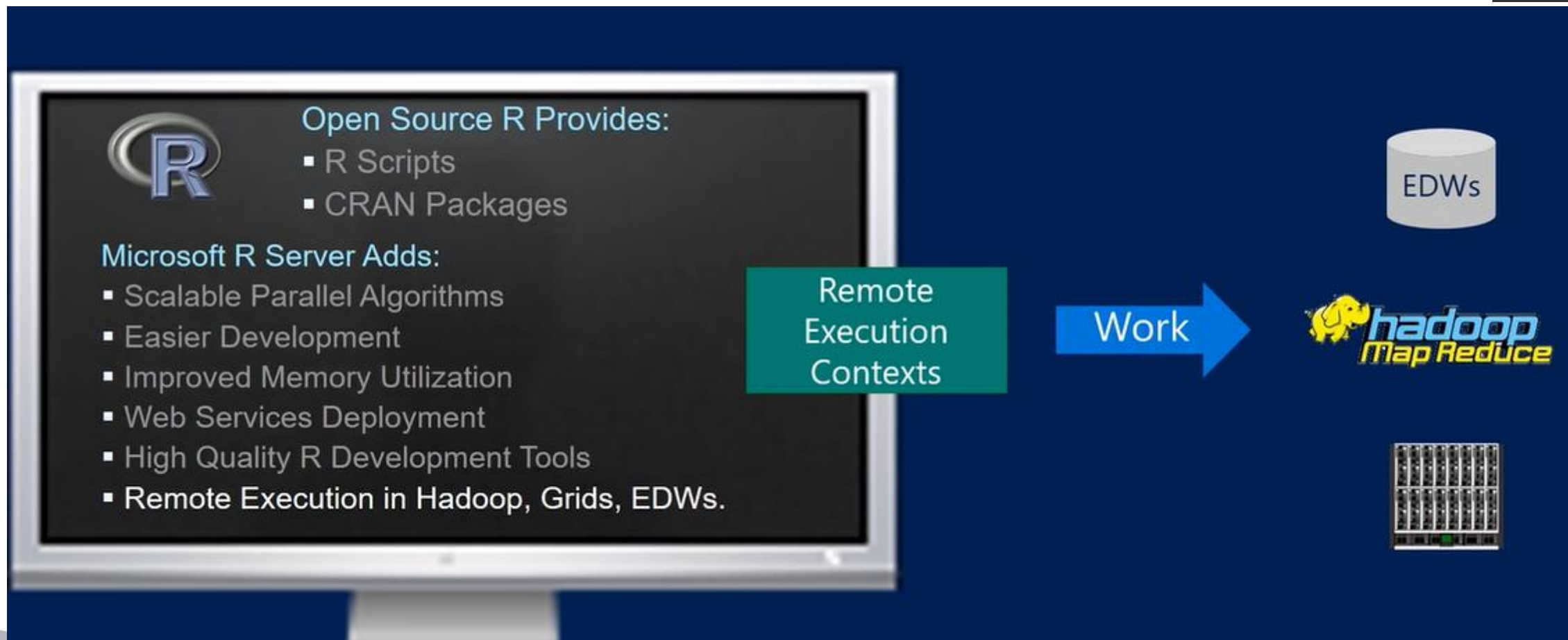
1. Starts A Master Process
2. Distribute Work
  1. Threading? Cores? Sockets? Nodes?
  2. Available RAM?
  3. Location of Data?
3. Master Tasks for Each Node



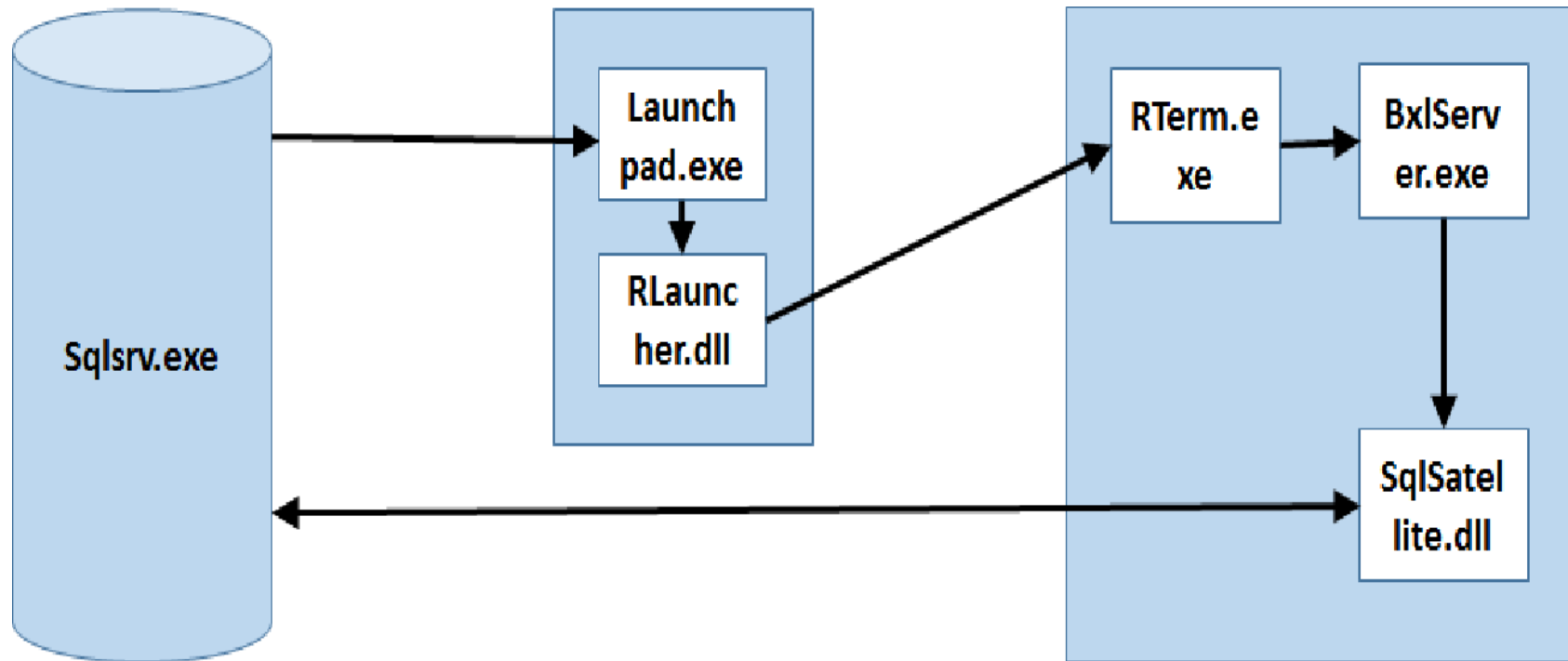
4. Master Initiates Distributed Work
  1. Hadoop Schedules Mapper for Each Split
  2. Algorithm Computes Intermediate Result
  3. Reducer Combines Intermediate Results
5. Master Process Evaluates Completion
6. Returns Consolidated Answer to Script



# For Client R / Server R – Remote execution



# Communication between R and SQL Server



# ScaleR algorithms



## Data Preparation

- 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

## Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: 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

## Variable Selection

- Stepwise Regression

## Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

## Cluster Analysis

- K-Means

## Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes



## Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms



# R code in SQL Server as T-SQL

```
EXECUTE sp_execute_external_script
 @language = N'R'
 ,@script = N'
 library(e1071);
 irismodel <-naiveBayes(iris_data[,1:4], iris_data[,5]);
 trained_model <- data.frame(payload = as.raw(serialize(irismodel, connection=NULL)));
 ,@input_data_1 = N'select "Sepal.Length", "Sepal.Width","Petal.Length","Petal.Width","Species" from iris_data'
 ,@input_data_1_name = N'iris_data'
 ,@output_data_1_name = N'trained_model'

WITH RESULT SETS ((model VARBINARY(MAX)));
```



# R code in SQL Server using Scale R algorithms



```
EXECUTE sp_execute_external_script
 @language = N'R'
 ,@script = N'require("RevoScaleR");
 irisLinMod <- rxLinMod(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width + Species,
 data = iris_rx_data);
 trained_model <- data.frame(payload = as.raw(serialize(irisLinMod, connection=NULL)));
 ,@input_data_1 = N'select "Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Species" from
 iris_rx_data'
 ,@input_data_1_name = N'iris_rx_data'
 ,@output_data_1_name = N'trained_model'

WITH result SETS ((model VARBINARY(MAX)));
```





# ScaleR Code

```

13 #####
14 #
15 #
16 # LOADING DATA (small sample)
17 # 178 MB
18 # 8.4Mio Rows
19 #
20 #####
21
22
23 ptm <- proc.time()
24 #inFile <- file.path(rxGetOption("sampleDataDir"), "AirlineDemoSmall.csv")
25 inFile <- file.path(rxGetOption("sampleDataDir"), "airsample.csv")
26 rxTextToxdf(inFile = inFile, outFile = "airline.xdf", stringsAsFactors = T, rowsPerRead = 200000, overwrite=TRUE)
27 proc.time() - ptm
28 # ~ 22 seconds!
29 # - 42 Chunks per 200.000 Rows; Total: 8.400.000 Rows
30
31 #####
32 # EXPLORING DATA (small sample)
33 #####
34
35
36 rxGetInfo(data="airline.xdf", getVarInfo = TRUE, numRows = 5)
37
38 #Histograms by day of week
39 ptm <- proc.time()
40 rxHistogram(~ ArrDelay|DayOfWeek, data = "airline.xdf")
41 proc.time() - ptm
42
43 #summary
44 rxSummary(~ ArrDelay, data = "airline.xdf")
45
46
47 rxSort(inData="airline.xdf", outFile = "sortFlights.xdf", sortByVars="ArrDelay", decreasing = TRUE, overwrite=TRUE)
48 # ~ 4 Seconds!
49 mostflights5 <- rxGetInfo(data = "sortFlights")
50 mostflights5
51 top5f <- as.data.frame(mostflights5[[5]])
52 topOA <- unique(as.vector(top5f$ArrDelay))
53 topOA
54
55
56
57 #####
58 # Linear Model with ReportProgress!
59 #####
60
61 # Linear Model using rxLinMod
62 sampleDataDir <- rxGetOption("sampleDataDir")
63 airlineDemoSmall <- file.path(sampleDataDir, "AirlineDemoSmall.xdf")
64
65 inLinMod <- rxLinMod(ArrDelay ~ CRSDepTime, data = airlineDemoSmall,

```



# Microsoft

# R code in SQL Server as T-SQL to generate graphs



```
DECLARE @RScript nvarchar(max)
DECLARE @SQLScript nvarchar(max)

SET @RScript = N'library(plotly)
library(ggplot2)
library(htmlwidgets)
#setwd("C:/DataTK/HTML")
image_file <- tempfile()
jpeg(filename = image_file, width = 500, height = 500)
df <- InputDataSet
d <- df[sample(nrow(df), 10),]
p <- plot_ly(d, x = OrderQty, y = DiscountPct, text = paste("OrderQty: ", OrderQty),
mode = "markers", color = OrderQty, size = OrderQty)
saveWidget(as.widget(p), "index.html")
OutputDataSet <- data.frame(data=readBin(file(image_file, "rb"), what=raw(), n=1e6))'

SET @SQLScript = N'SELECT
 ps.[Name]
 ,AVG(sod.[OrderQty]) AS OrderQty
 ,so.[DiscountPct]
 ,pc.name AS Category
FROM Adventureworks.[Sales].[SalesOrderDetail] sod
INNER JOIN Adventureworks.[Sales].[SpecialOffer] so
ON so.[SpecialOfferID] = sod.[SpecialOfferID]
INNER JOIN Adventureworks.[Production].[Product] p|
ON p.[ProductID] = sod.[ProductID]
INNER JOIN Adventureworks.[Production].[ProductSubcategory] ps
ON ps.[ProductSubcategoryID] = p.ProductSubcategoryID
INNER JOIN Adventureworks.[Production].[ProductCategory] pc
ON pc.ProductCategoryID = ps.ProductCategoryID
GROUP BY ps.[Name],so.[DiscountPct],pc.name'

EXECUTE sp_execute_external_script
@language = N'R',
@script = @RScript,
@input_data_1 = @SQLScript
WITH RESULT SETS ((Plot varbinary(max)))
```





# Benefits of R integration

- Based on Open source R
- Different versions available (Open, Client and Server)
- Distributed workloads, multi-threading and parallelization
- Interoperable (Windows, Linux, MacOS) with different flavors (Hadoop, Teradata, HDInsight)
- Faster model prediction and model deployment
- No „in-memory“ constraints, less data movement, less bottlenecks in performance, no data size limitations
- Hybrid topologies, agile development, stable platform for data operationalization, investment protection (SLA, Terms and agreements)
- R Code is available in SSMS environment
- Community and commercial support
- R Language is growing in popularity





Bunch of demos



SQL Server vNext (CTP1 and above)





# Microsoft R Server 9.0 (MRS9)

- SQL Server vNext (CTP1)
- R - 3.3.2
- RevoScaleR (9.0.3)
- MicrosoftML (1.0.0)



# Quick recap: RevoScaleR (9.0.1)

## Importing functions and computation context

rxImport  
rxDataStep  
rxGetInfo  
rxSetInfo  
rxGetVarInfo  
rxSetVarInfo  
rxGetVarNames  
rxCreateCollInfo  
rxCompressXdf  
RxXdfData  
RxTextData  
RxOdbcData  
RxSqlServerData  
rxOpen  
rxClose  
rxReadNext  
rxSetFileSystem  
rxGetFileSystem  
rxNativeFileSystem  
rxSetComputeContext  
rxGetComputeContext  
RxHadoopMR  
RxInSqlServer  
RxComputeContext  
RxLocalSeq  
RxLocalParallel  
RxForeachDoPar  
rxInstalledPackages  
rxFindPackage

## Data Manipulation

xDataStep  
rxFactors  
rxGetFuzzyDist  
rxGetFuzzyKeys  
rxSplit  
rxSort  
rxMerge  
rxExecuteSQLDDL

## Data Visualization

rxHistogram  
rxLinePlot  
rxLorenz  
rxRocCurve

## Descriptive /cross-tab Statistics

rxQuantile  
rxSummary  
rxCrossTabs  
rxCube  
rxMarginals  
as.xtabs  
rxChiSquaredTest  
rxFisherTest  
rxKendallCor  
rxPairwiseCrossTab  
rxRiskRatio  
rxOddsRatio

## Analysis and Predictive statistics

**rxLinMod**  
**rxLogit**  
**rxGlm**  
rxCovCor  
**rxDTree**  
**rxBTrees**  
**rxDForest**  
rxPredict  
rxKmeans  
**rxNaiveBayes**  
rxCov  
rxCor  
rxSSCP  
rxRoc

Shorter list of functions





# Quick recap: RevoScaleR (9.0.1)

|                                   | Task                     |                              |                |                | Scalability |         |                  | Description |
|-----------------------------------|--------------------------|------------------------------|----------------|----------------|-------------|---------|------------------|-------------|
|                                   | Predict categories       |                              | Predict values |                |             |         |                  |             |
| RevoScaleR<br>functions / learner | Binary<br>classification | Multiclass<br>classification | Regression     | Other          | rows        | columns | CPU /<br>threads |             |
| rxLinMod                          |                          |                              | Yes            |                | 100 Mil.    | 10 K    | Multi            |             |
| rxLogit                           | yes                      |                              |                |                | 100 Mil.    | 10 K    | Multi            |             |
| rxGlm                             | yes                      |                              | Yes            |                | 10 Mil      | 5 K     | Multi            |             |
| rxDTree                           | Yes                      | Yes                          | Yes            |                | 100 Mil.    |         | Multi            |             |
| rxBTrees                          | Yes                      | yes                          | Yes            |                | 100K        | 1K      | Multi            |             |
| rxDForest                         | Yes                      | Yes                          | Yes            |                |             |         | Multi            |             |
| rxKmeans                          |                          |                              |                | Classification |             | <10 K   |                  |             |
| rxNaiveBayes                      | Yes                      | Yes??                        |                |                | 100Mil      |         | Multi            |             |

All functions / learners work with XDF data formats



# MicrosoftML (1.0.0)



- **New fast and accurate learners** (Sentiment analysis, Customer Churn, Loadn risk prediction, demand prediction)
- **Text Classification** (Sentiment Analysis, Classification of Support ticket, Complaint book,. etc)
- **DNN with GPU Accelleration** (Retail image matching, medical image classification, metal/iron/steel industry control check (with live stream))
- **High-Dimensional Categorical data** (regression predictions with a lot of productIDs, customerIDs, Web Analytics (click through predictions))



# Algorithms in MicrosoftML

|                      | Task                  |                           |                |                   | Scalability |         |               | Description                                                      |
|----------------------|-----------------------|---------------------------|----------------|-------------------|-------------|---------|---------------|------------------------------------------------------------------|
|                      | Predict categories    |                           | Predict values |                   |             |         |               |                                                                  |
| MicrosoftML Learner  | Binary classification | Multiclass classification | Regression     | Other             | rows        | columns | CPU / threads |                                                                  |
| rxFastLinear         | Yes                   |                           | Yes            |                   | 1 Bil.      | 1 Bil.  | Multi         | Fast Linear (SDCA) with L1 & L2                                  |
| rxLogisticRegression | Yes                   | Yes                       |                |                   |             | 100 Mil |               | Logistic regression with L1 & L2                                 |
| rxNeuralNet          | Yes                   | Yse                       | Yes            |                   | unlimited   | 10 Mil  | Multi / GPU   | Neural Network / GPU-accelerated NET# DNN with convolutions      |
| rxFastTree           | Yes                   |                           | Yes            |                   |             | 50 K    | Multi         | Boosted Decision Tree                                            |
| rxFastForest         | Yes                   |                           | Yes            |                   |             | 50 K    | Multi         | Random Forest                                                    |
| rxOneClassSvm        |                       |                           |                | Anomaly/Reduction |             | 1 K     | Single        | Anomaly detection / reduction / unbalanced binary classification |

All learners work with XDF data formats



# Questions?



<http://tomaztsql.wordpress.com>



tomaz.kastrun@gmail.com



@tomaz\_tsq1



/in/tomaztsql



<http://github.com/tomaztk>



<https://mvp.microsoft.com/PublicProfile/5002196>