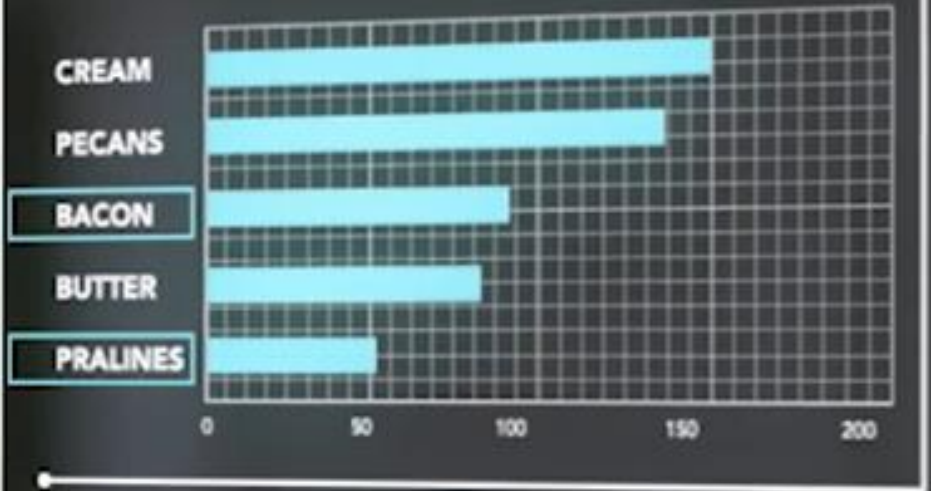


**BEST SELLER:
PECANS & CREAM**

▣ SOCIAL AFFINITY SEARCH



SENTIMENT ANALYSIS: BACON + PRALINES



Microsoft R Server Overview

Slalom Consulting
Instructors & proctors:
Dan Tetrick, Ben Ahlvin, Janet Guerrero

Day One Modules

- The Microsoft R Data Stack
- The Core Principles of R
- Functional Programming in R
- Deep-dive into the dplyr package for data manipulation
- The dplyrXdf Package: using dplyr syntax with xdf

Day Two Modules

- Modeling and Scoring with Microsoft R Server
- Parallel Computing
- Azure Portal: provisioning a Data Science Virtual Machine
- HDInsights Insights
- SQL Server and R Services

Quick Poll

(1) What is your level of R experience?

Low/Medium/High

(2) What do you hope to get out of the class?

What is



Language Platform

- A statistics programming language
- A data visualization tool
- Open source
- Focus on statistics and machine learning
- Single-threaded
- Data stored in memory

Community

- 2.5+M users
- Taught in most universities
- New and recent grad's use it
- Thriving user groups worldwide


Ecosystem

- 10,000+ free algorithms in CRAN
- Scalable to big data
- Rich application & platform integration

CRAN: Comprehensive R Archive Network

CRAN Task Views


CRAN Task Views are guides to R packages and functions useful for solving specific problems. Many R packages have a Task View associated with them. As an effort to make them more visible, we have created a Task View page. The Task View page is the best place to go to find a package that can help you solve a problem. The Task View page is the best place to go to find a package that can help you solve a problem.



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...


[\[more\]](#)



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...


[\[more\]](#)



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...

[\[more\]](#)



Psychometric Models and Methods

Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked...

[\[more\]](#)

Bayesian Inference

Clinical Trial Design, Monitoring, and Analysis

Cluster Analysis & Finite Mixture Models

Probability Distributions

Computational Economics

Design of Experiments (DoE) & Analysis of Experimental Data

Empirical Finance

Statistical Genetics

Natural Language Processing

Analysis of Pharmacokinetic Data

Official Statistics & Survey Methodology

Survival Analysis

Time Series Analysis

Statistical Methods

Machine Learning & Statistical Learning

Graphic Displays & Dynamic Graphics, Graphic Devices & Visualization

Medical Image Analysis

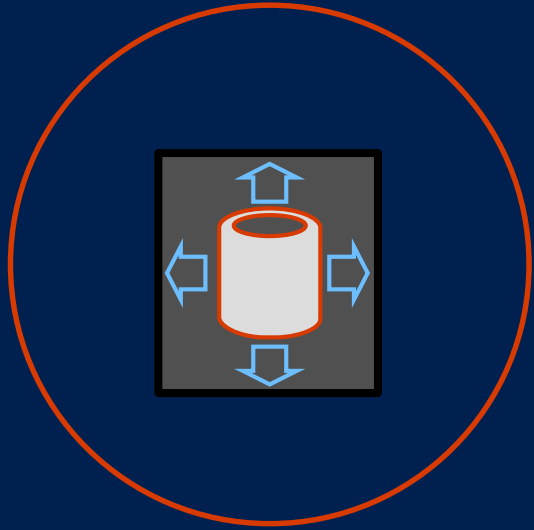
Psychometric Models and Methods

Statistics for the Social Sciences

gGraphical Models

In addition to CRAN, Bioconductor, GitHub, others distribute R packages

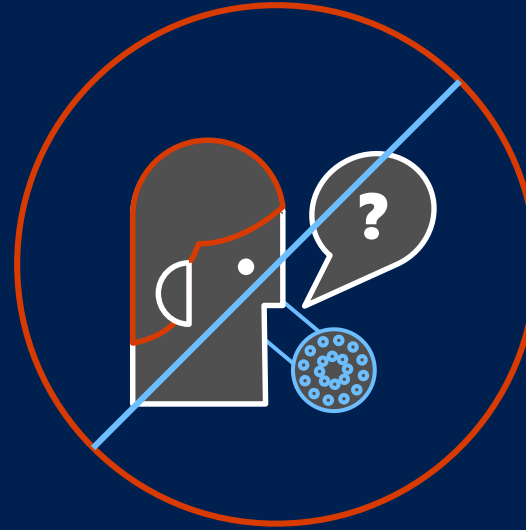
Challenges posed by open source R



Limited
Data
Scale



Inadequate
Modeling
Performance



Lack of
Commercial
Support



Complex
Deployment
Processes

Microsoft R Server

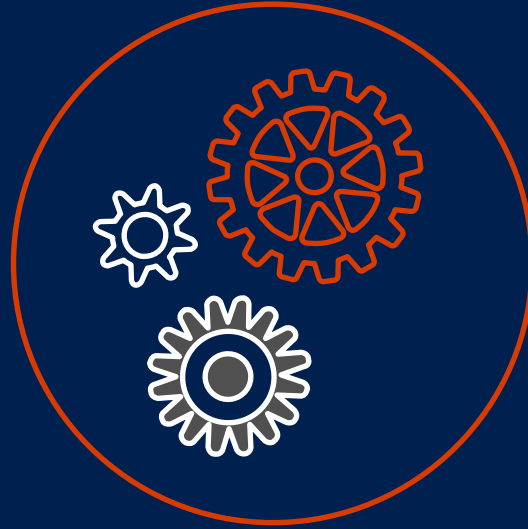
MRS extends open-source R to allow:

- Multi-threading
 - Matrix operations, linear algebra, and many other math operations run on all available cores
- Parallel processing
 - ScaleR functions utilize all available resources, local or distributed
- On-disk data storage
 - RAM limitations lifted – Break Through Your Memory Barrier!

R from Microsoft brings



Peace of
mind



Efficiency



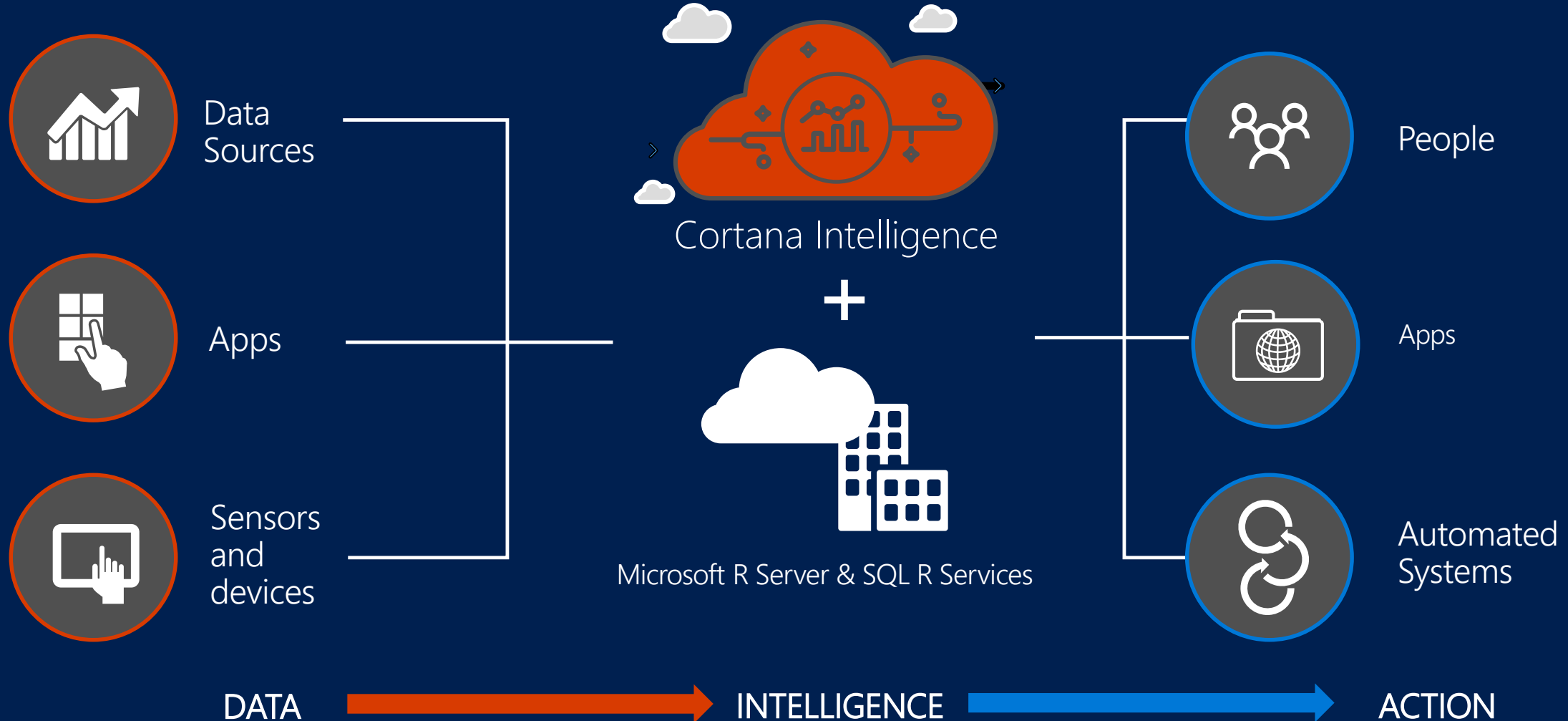
Speed and
scalability



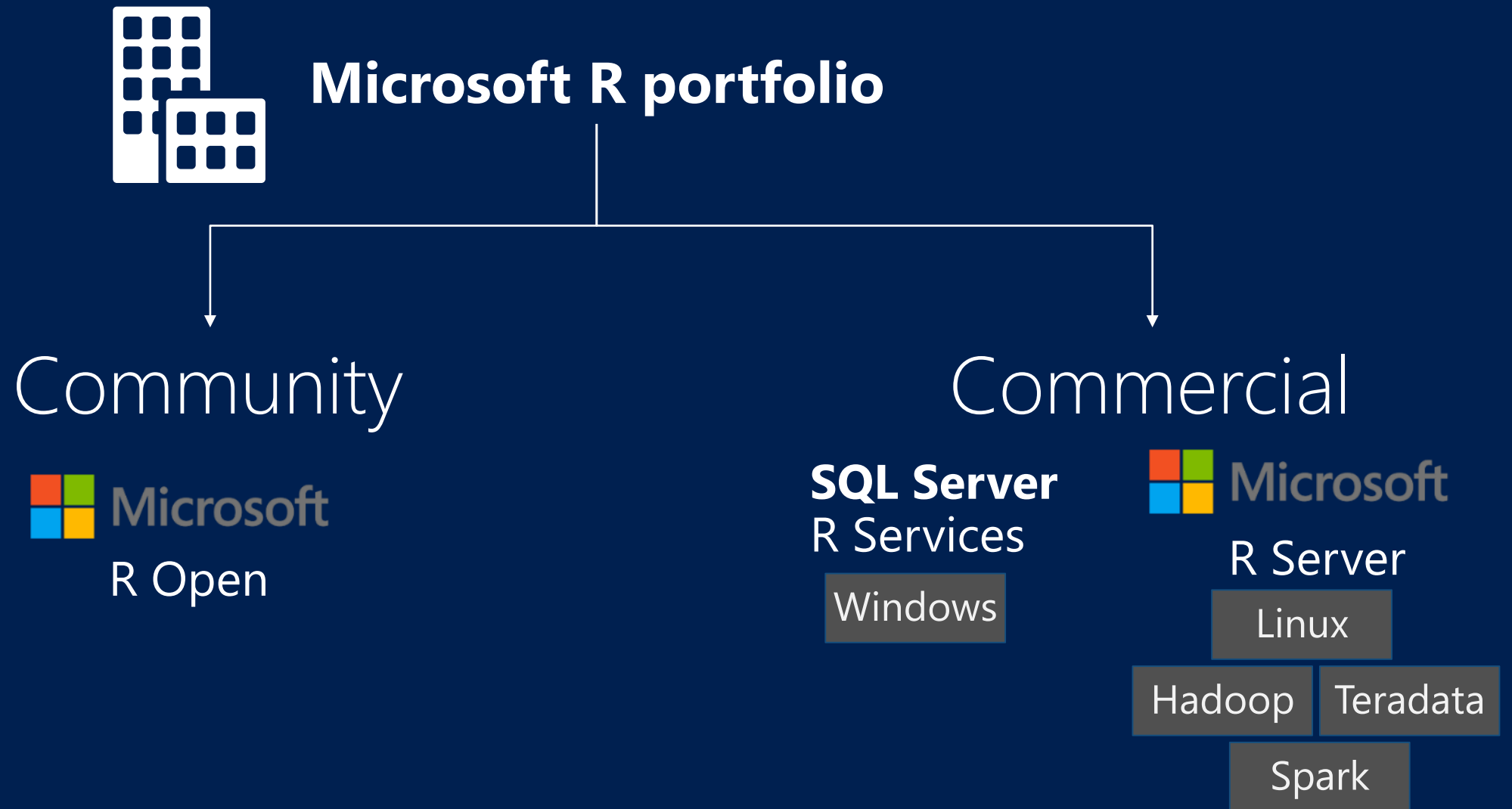
Flexibility
and agility

Microsoft R Server family

From Data To Action On Premises



Microsoft R portfolio

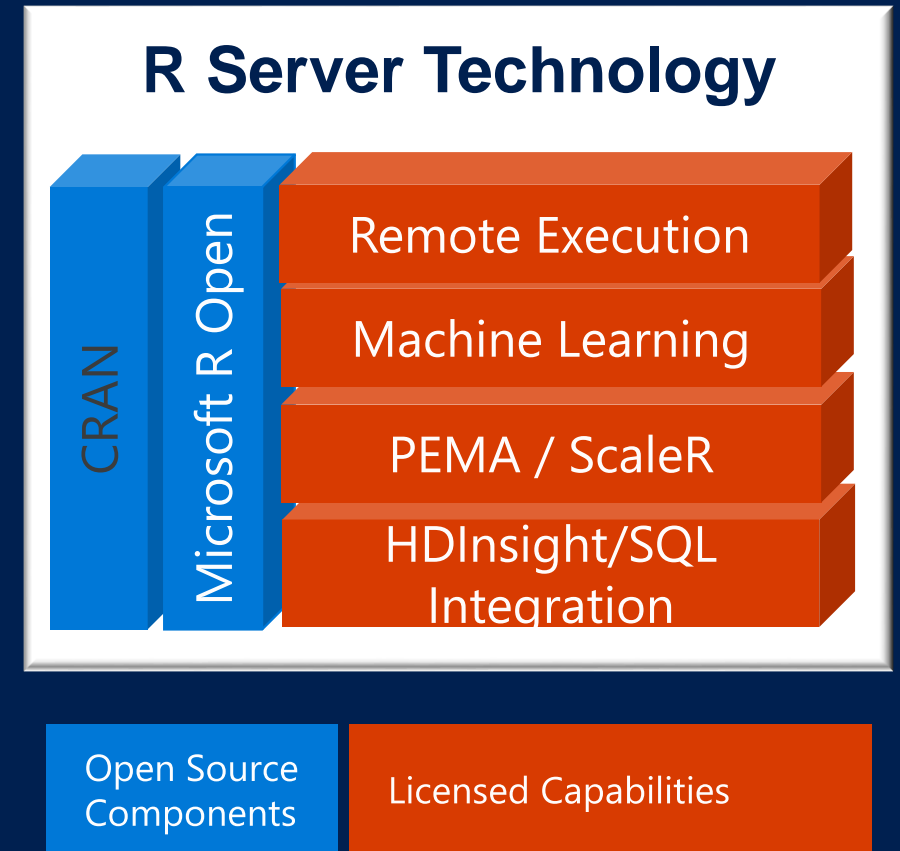


Linux, Windows, Hadoop & Teradata

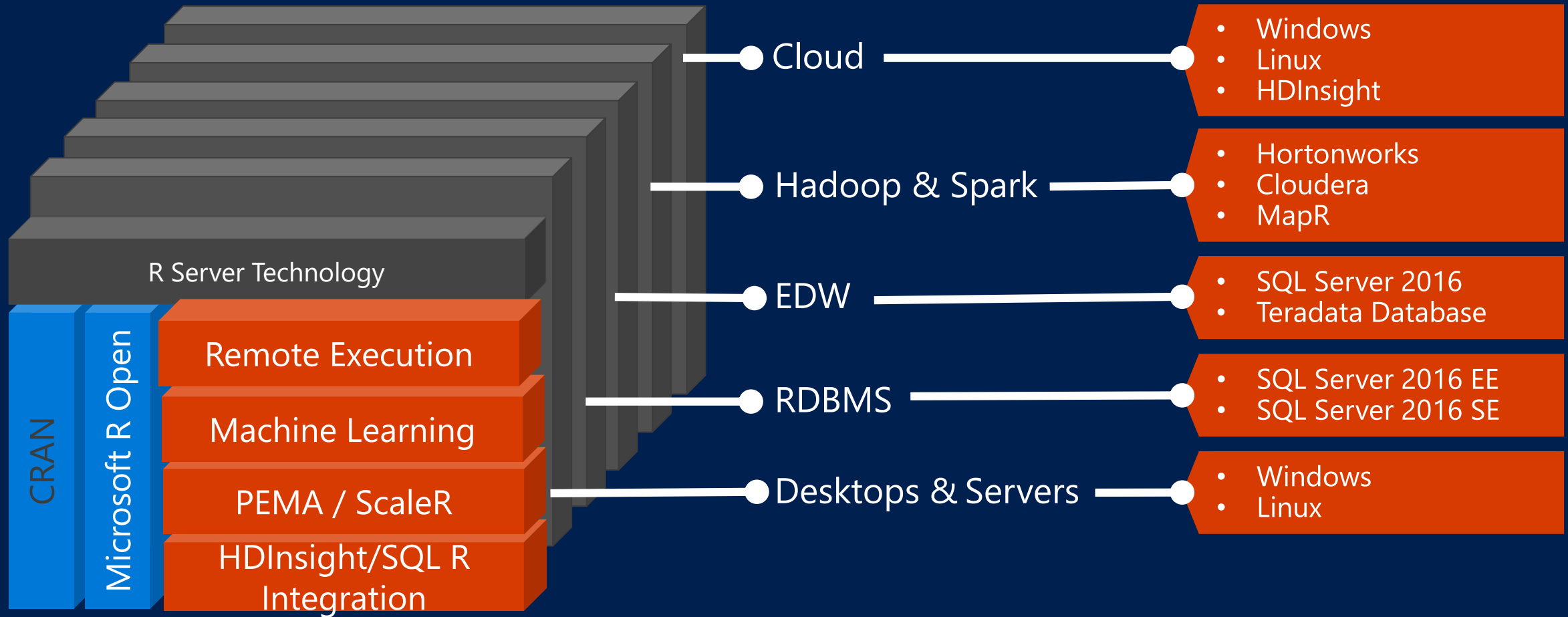
Introducing Microsoft R Server

High-performance, Scalable R

- 100% open source R
- CRAN, Bioconductor, MRAN, GitHub compatibility
- Big-data connectivity
- Scalable analytics
- Multi-platform
- In-database, in-cluster scalability
- Choice of IDE

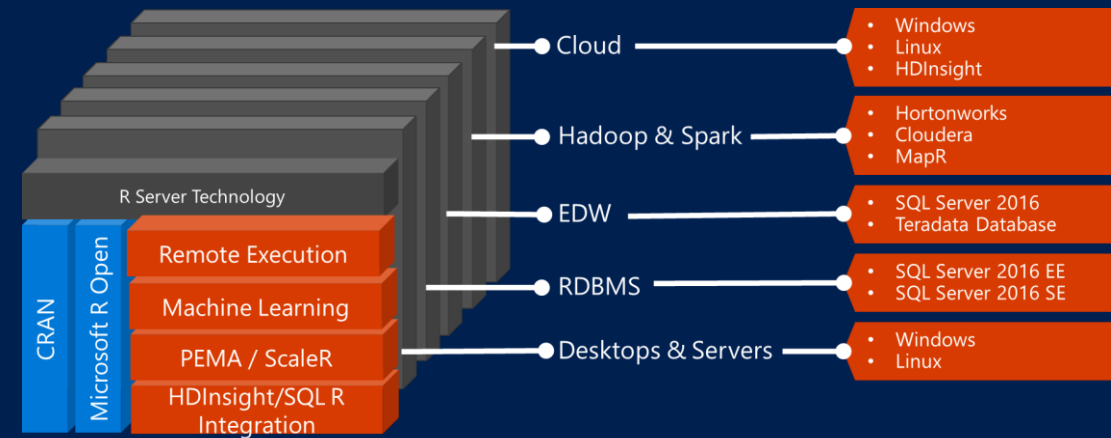


Portability & investment assurance



Write Once – Deploy Anywhere

MRS in Different Contexts



- On a workstation, that means:
 - All available cores will be used for math operations and parallel processes
 - Hard drive capacity sets limit for data size, not RAM
- On a cluster:
 - Parallel utilization of all available nodes
 - Distributed file systems like HDFS greatly expand possible data sizes

Available Algorithms

- Linear regression (rxLinMod)
- Generalized linear models (rxLogit, rxGLM)
- Decision trees (rxDTree)
- Gradient boosted decision trees (rxBTree)
- Decision forests (rxDForest)
- K-means (rxKmeans)
- Naïve Bayes (rxNaiveBayes)

Note: models available in open-source R packages won't be made parallel automatically



Parallelized, Remote Execution Algorithms

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

Predictive Models

Sum of Squares (cross product matrix for set variables)

Quantiles (approx.)

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

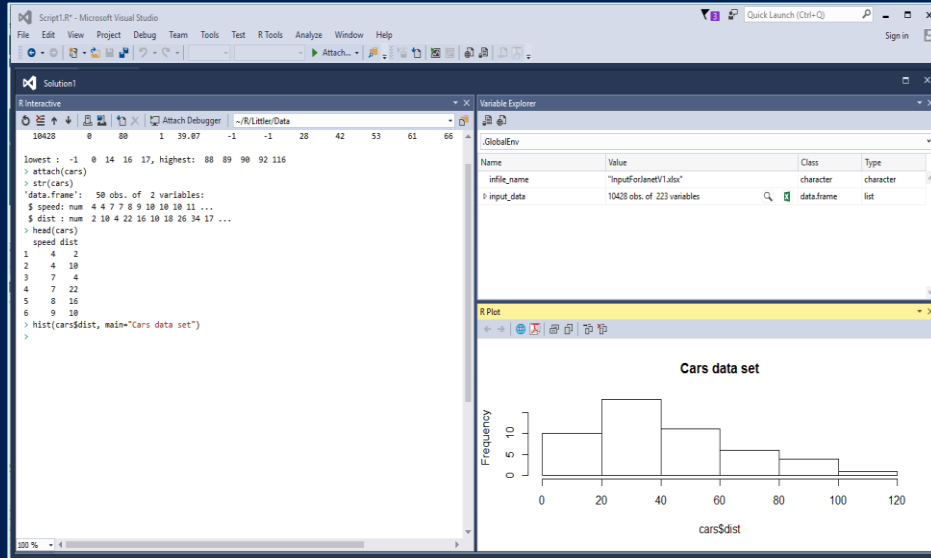
PEMA-R API Custom Algorithms



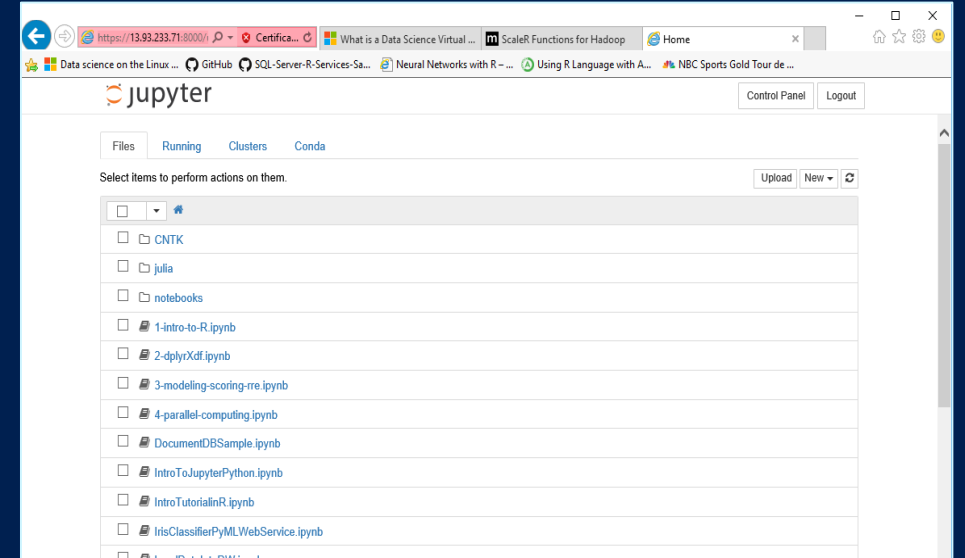
Summary: use MRS when...

- Working with data too big to fit into memory
- Building models that take too long to run
- Working with clusters and distributed file systems

There are several R Clients

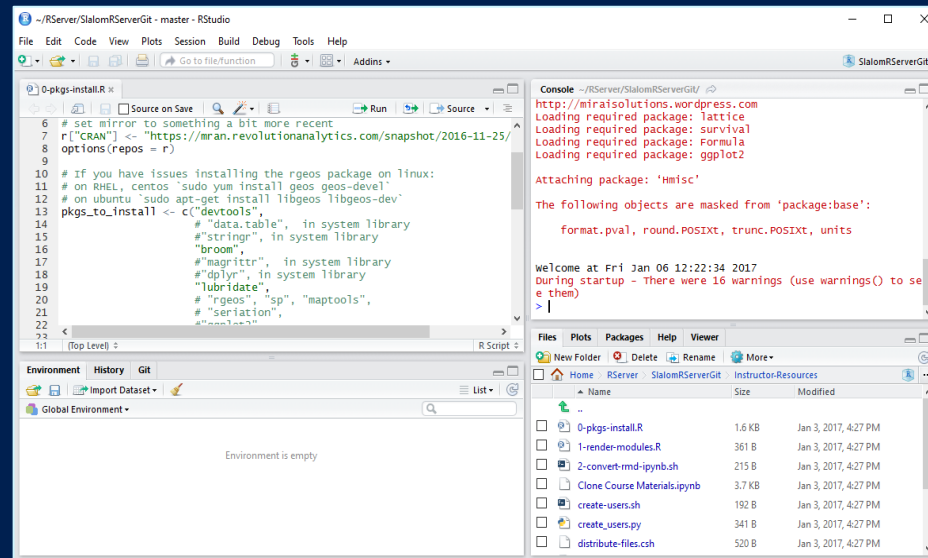


Microsoft R Client (VS)



Jupyter Notebooks

R Studio



Appendix