Microsoft R server

Stefan Cronjaeger

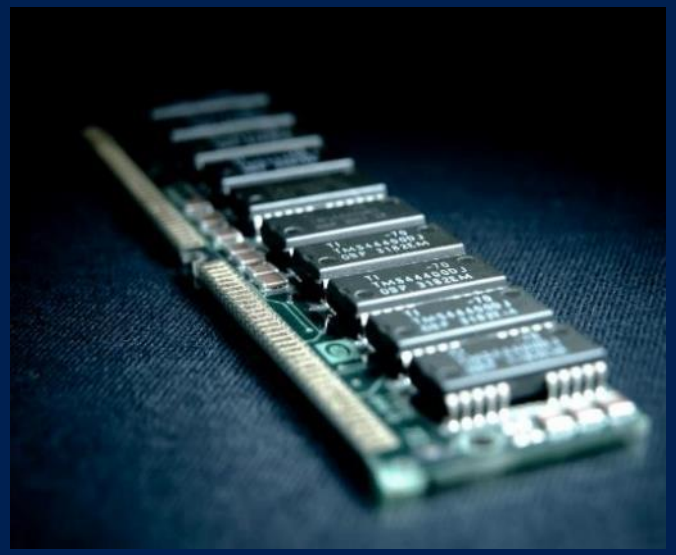
Technical Solution Specialist Advanced Analytics
Global Blackbelt – Germany
scronj@microsoft.com
+49 151 4406 3425

Microsoft

Topics

- Microsoft R Server (and demo)
- R on Hadoop
- R on SQL server
- Operationalizing R

Enterprise use of open source R



R needs data in memory to start a computation*

R is mostly single threaded

R requires skilled resource to scale out computations across a cluster and needs re-coding for R map-reduce in Hadoop

Open source R is supported by the community

CRAN, MRO, MRS Comparison

		**Microsoft R Open**	**Microsoft R Server**
Datasize	In-memory	In-memory	**In-Memory or Disk Based**
Speed of Analysis	Single threaded	Multi-threaded	**Multi-threaded, parallel processing 1:N servers**
Support	Community	Community	**Community + Commercial**
Analytic Breadth & Depth	7500+ innovative analytic packages	7500+ innovative analytic packages	**7500+ innovative packages + commercial parallel high-speed functions**
Licence	Open Source	Open Source	**Commercial license. Supported release with indemnity**

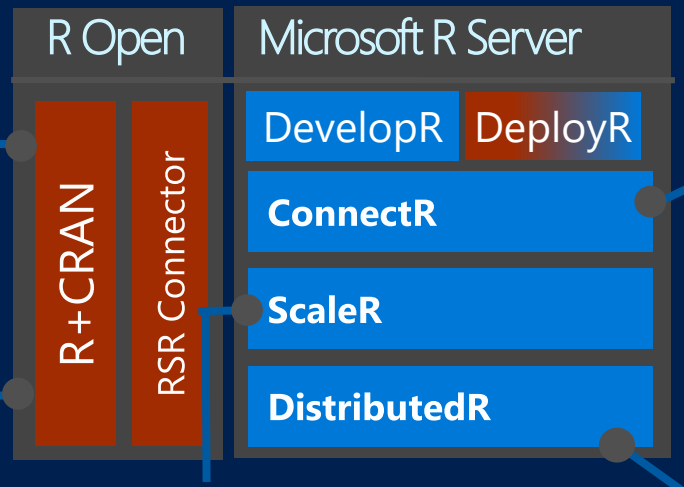
The Microsoft R Server Platform

R+CRAN

- Open source R interpreter
  - R 3.1.2
- Freely-available huge range of R algorithms
- 100% Compatible with existing R scripts, functions and packages

RevoR

- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math library to speed up linear algebra functions



ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics

ConnectR

- High-speed & direct connectors

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Teradata Database & Aster
- ODBC

DistributedR

- Distributed computing framework
- Delivers cross-platform portability

Scale R – Parallelized Algorithms & Functions

Data Preparation

- 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

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

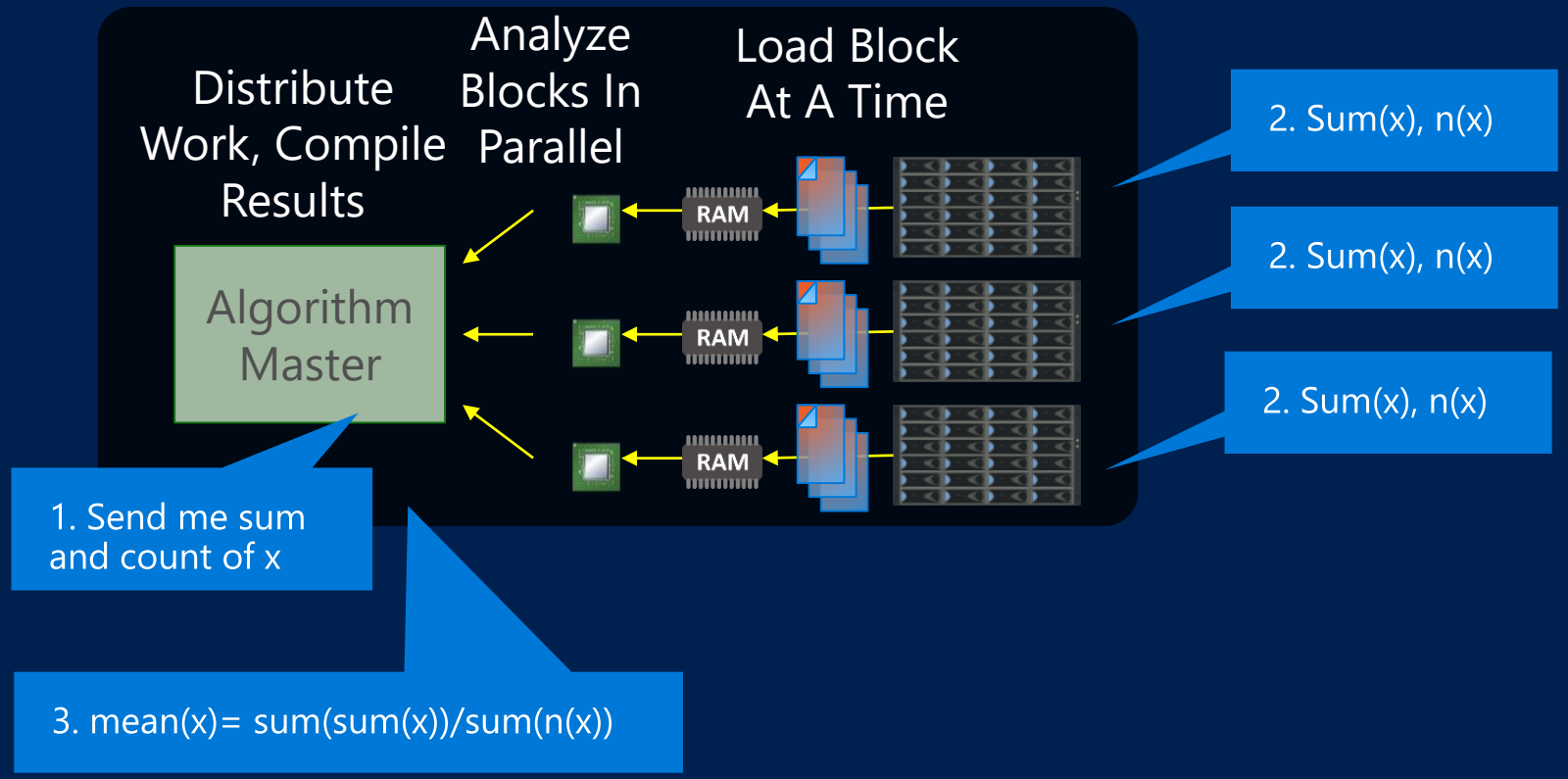
Classification

- Decision Trees
- Decision Forest
- Gradient Boosted Decision Trees
- Naïve Bayes

Combination

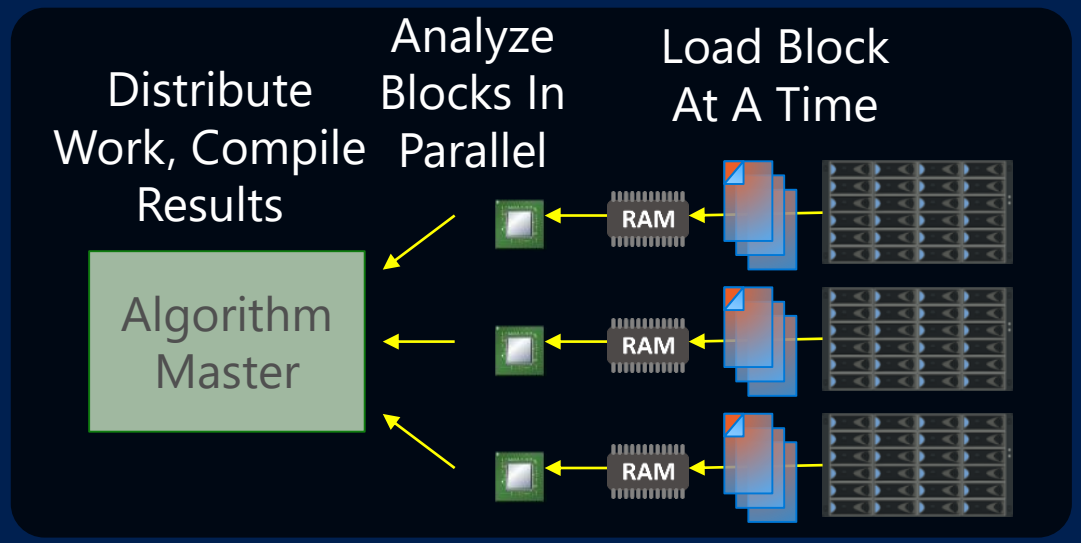
- rxDataStep
- rxExec
- PEMA API

Parallel Work Simple Example: Mean



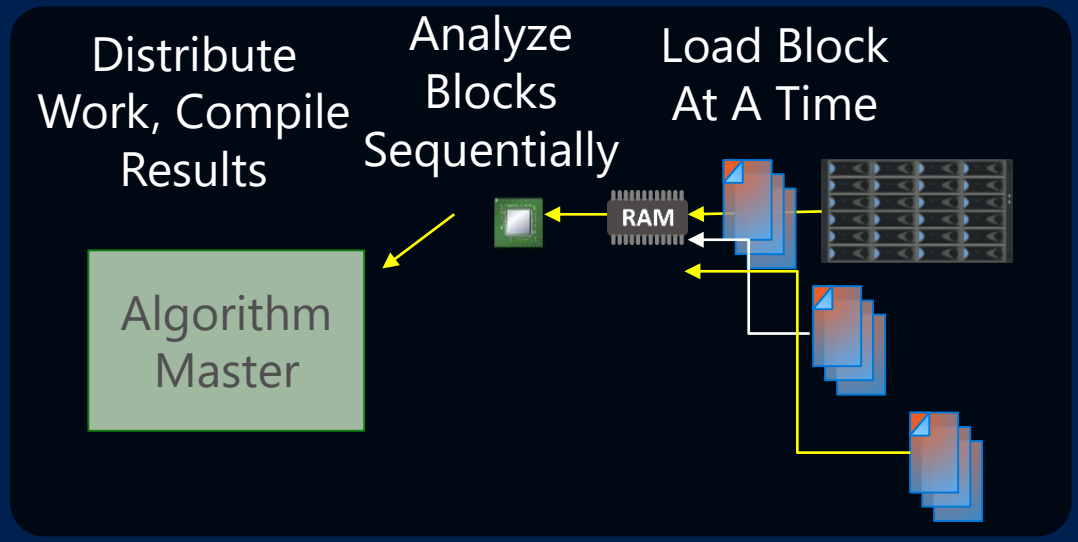
Not every algorithm works in parallel
Often there are several steps involved

Parallelization or Sequential Execution



Parallel:

- Large number of parallel workers
- Fast handling of Big Data



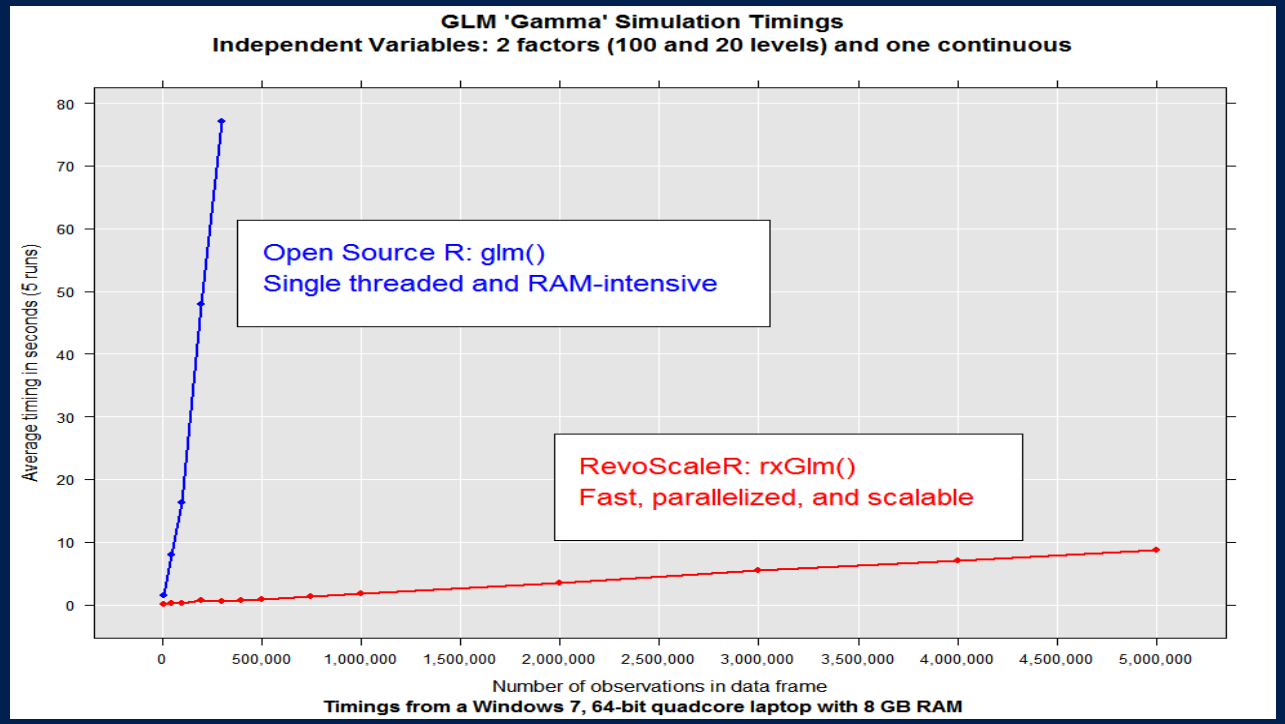
Sequential:

- Just one block of data in RAM
- Handling of Big Data with moderate resources

Demo Microsoft R Server

ScaleR - Performance comparison

Microsoft R Server has no data size limits in relation to size of available RAM



File Name	Compressed File Size (MB)	No. Rows	Open Source R (secs)	Revolution R (secs)
Tiny	0.3	1,235	0.00	0.05
V. Small	0.4	12,353	0.21	0.05
Small	1.3	123,534	0.03	0.03
Medium	10.7	1,235,349	1.94	0.08
Large	104.5	12,353,496	60.69	0.42
Big (full)	12,960.0	123,534,969	Memory!	4.89
V.Big	25,919.7	247,069,938	Memory!	9.49
Huge	51,840.2	494,139,876	Memory!	18.92

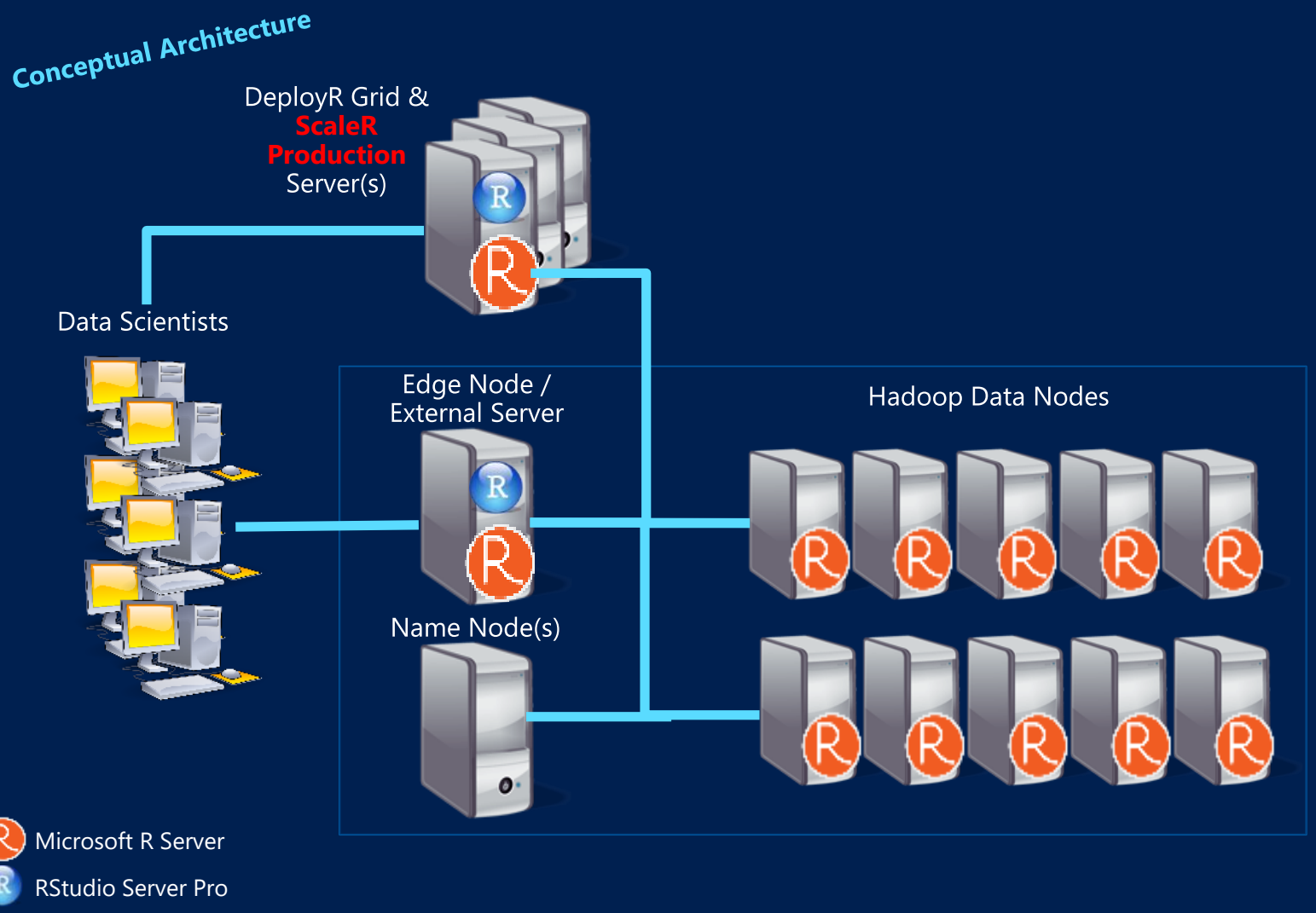
- US flight data for 20 years
- Linear Regression on Arrival Delay
- Run on 4 core laptop, 16GB RAM and 500GB SSD

Microsoft R Server

Hadoop

Microsoft

MRS and Hadoop Architecture options



Microsoft R model run options:

1. **Copy** HDFS data to Edge Node / External Server Linux file system. Use "**Local Parallel**" compute context on that server to run model
2. **Stream** data from HDFS to Edge Node / External Server. Use "**Local Parallel**" compute context on that server to run model and discard data
3. **Send** the Microsoft R model script to run in every data node and return model output object to the Edge Node / External Server

Write Once Deploy Anywhere

ScaleR functions can run in-Hadoop, in-Spark or in-Database without any functional R recoding

And Spark

And Teradata

Local Parallel – Linux or Windows

```
# SETUP LINUX ENVIRONMENT VARIABLES
rxSetComputeContext("localpar")

# CREATE LINUX, DIRECTORY AND FILE OBJECTS
linuxFS <- RxNativeFileSystem()

AirlineDataSet <- RxXdfData("AirlineDemoSmall.xdf", fileSystem = linuxFS)
```

In – Hadoop

```
### SETUP HADOOP ENVIRONMENT VARIABLES
myHadoopCluster <- RxHadoopMR()

### HADOOP COMPUTE CONTEXT USING HDFS
rxSetComputeContext(myHadoopCluster)

### CREATE HDFS, DIRECTORY AND FILE OBJECTS
hdfsFS <- RxHdfsFileSystem()
 AirlineDataSet <-
 RxXdfData("AirlineDemoSmall.xdf",
 fileSystem = hdfsFS
```

SQL Server

```
# SETUP SQLSERVER ENVIRONMENT VARIABLES
mySqlServer <- RxInSqlServer()

# SQL SERVER COMPUTE CONTEXT AND TABLE REF
rxSetComputeContext(mySqlServer)

AirlineDataSet <-
RxSqlServerData(table="AirlineDemoSmall")
```

R script – does not need to change to run across different platforms

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
   rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
   rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
   hdfsXdfArrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data =
   AirlineDataSet)
   plot(hdfsXdfArrLateLinMod$coefficients)
```

Microsoft R Server

SQL Server R Services

Microsoft

Why In-Database Analytics with SQL 2016 & R?

Leverage Full Capability of R:

- Rich Statistical, Visualization & Predictive Analytics
- A Large and Growing Skill Base

… including Microsoft R Servers Big Data Capabilities:

- Scalable Computation
- Scalable Data Size

… all Running In-Database:

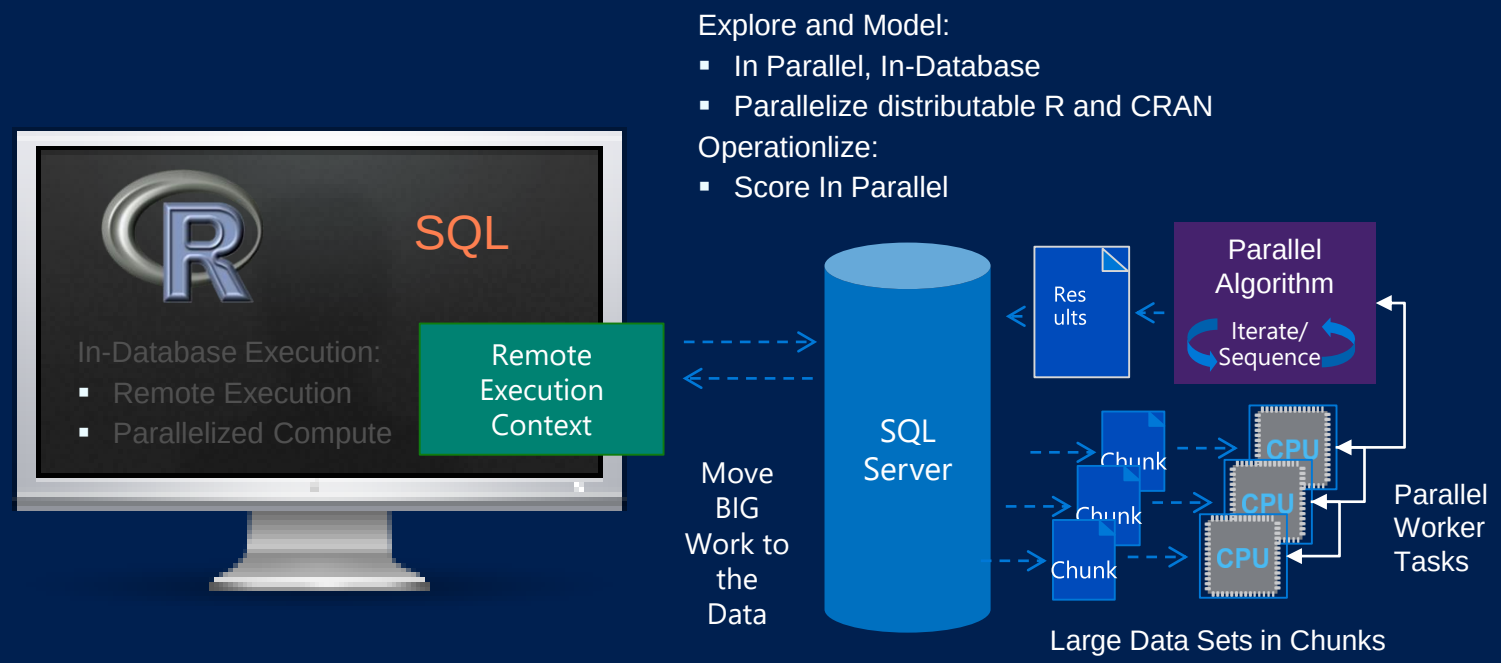
- Divide Work Between Data Scientists and Data Engineers
- Reduce Data Duplication and Data Movement

… While Protecting Information:

- Eliminate Data Movement & Unnecessary Copying
- Leverage Database Data Protections
- Leverage Database Tools for Backup, Scheduling, …

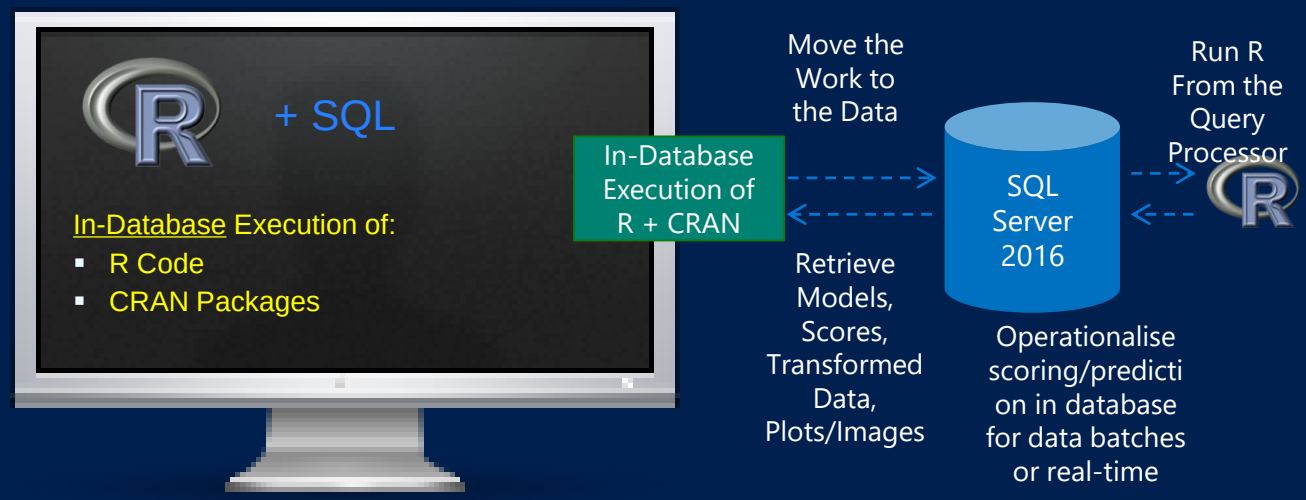
Run Parallel Algorithms in Database from an R client

- IDE for R
  - For Data Scientists
  - R coding
- SQL statements for data access
- SQL compute context
- Know-how:
  - R developer
  - Data science



Run R In-Database from TSQL

- IDE for SQL
  - For SQL developers, DB admins
- R code in stored procedures
- Know-how:
  - SQL developer
  - Operationalizing of SQL
  - Back-up, security, access control



Microsoft R Server

Operationalizing R based Analytics

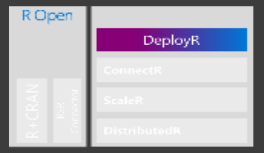
Microsoft

DeployR: Framework for R as a service for BI / web apps

Desktop Applications | Mobile Apps | Web Applications | Real-Time Applications

Business User Layer

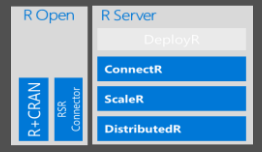
DeployR Web Services Layer



API Client Libraries

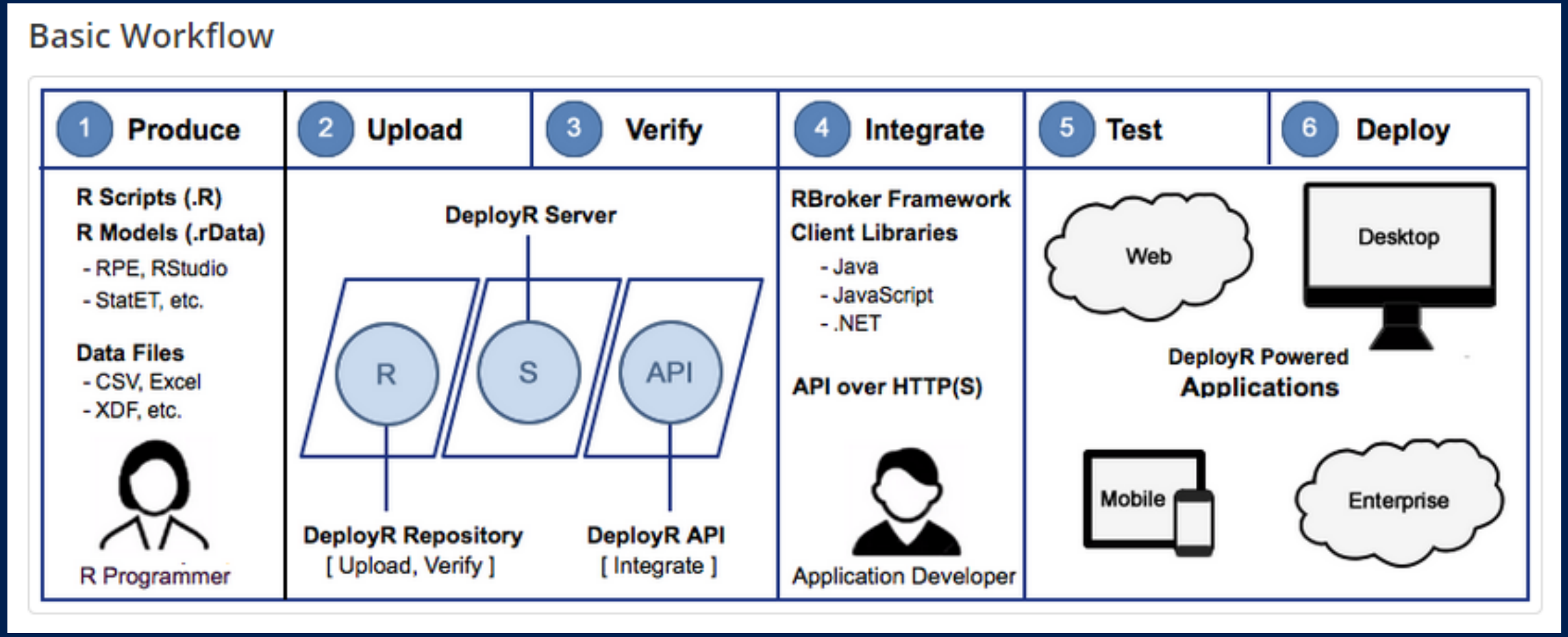
R Model Repository | **Enterprise Security** | **R Session Management** | **Resource Management**

R / Statistical Modeling Expert Layer



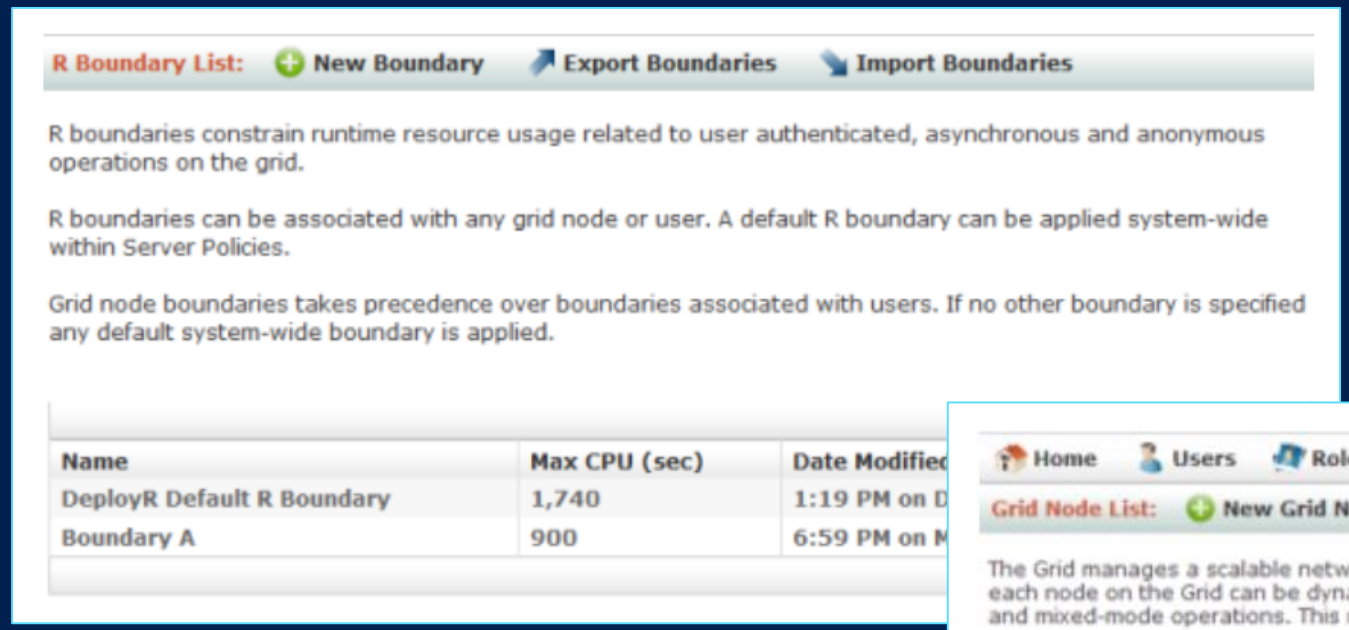
Revolution Scale R engine +
Open Source R with access to 7000+ packages

DeployR Workflow



DeployR – Administration

The creation and management of R runtime resource usage



The monitoring of events on the grid and server

Azure Munich Meetup

München, Deutschland
Gegründet 23. Jul 2016

Über uns...

Freunde einladen

Mitglieder	161
Anstehende	2

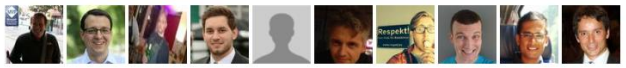
Herzlich willkommen!

Schlage ein neues Meetup vor

Anstehende (2) Vergangene Kalender

Azure IoT Development and Azure IoT Security

Microsoft Deutschland GmbH
Walter-Gropius-Str. 5, München (Karte)



TBD weitere Infos

Veranstaltet von: Christian Waha (Organisator)

Di, 22. Nov
19:00
RSVP
43 gehen hin
0 Kommentare

IoT und Industrie 4.0: Von Buzzwords zu neuen Geschäftsmodellen

Microsoft Deutschland GmbH
Walter-Gropius-Str. 5, München (Karte)

Di, 13. Dez
19:00

Was gibt es Neues

NEUES MITGLIED
Martin Meixger macht mit
vor 2 Stunden

NEUES MITGLIED
Christian Straus macht mit
Gestern

NEUES RSVP
Marcus Franzen hat Ja für IoT und Industrie 4.0: Von Buzzwords zu neuen Geschäftsmodellen geantwortet
Vor 2 Tagen

NEUES RSVP
Marcus Franzen hat Ja für Azure IoT

all azure communities?

http://aka.ms/azure-meetups