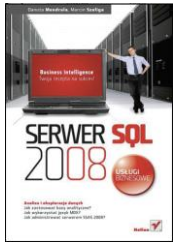
od STUDENTA do SPECJALISTY BUSINESS INTELLIGENCE konferencja

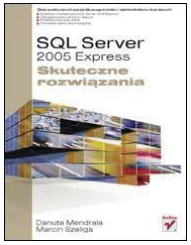
Revolutionary R integration with SQL Server 2016

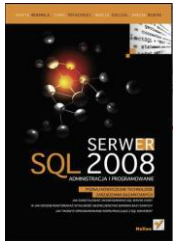
Marcin Szeliga

Marcin Szeliga

- Data Philosopher
- 20 years of experience with SQL Server
- Data Platform MVP & MCT
- MCSE: Data Platform & Business Intelligence
- MCSD: Azure & Microsoft.NET Solutions Architect
- marcin@sqlexpert.pl





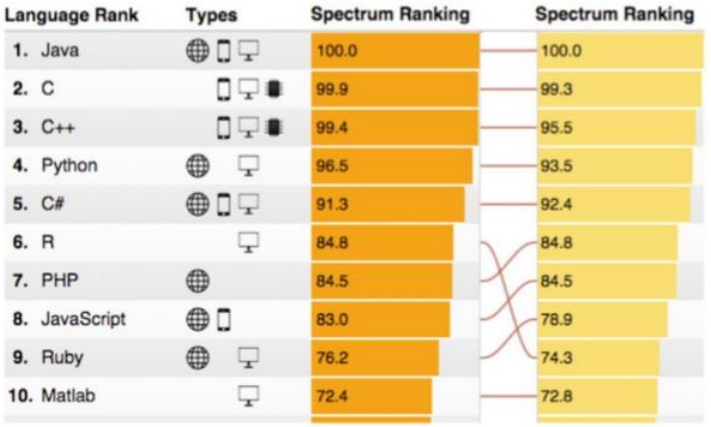


Agenda

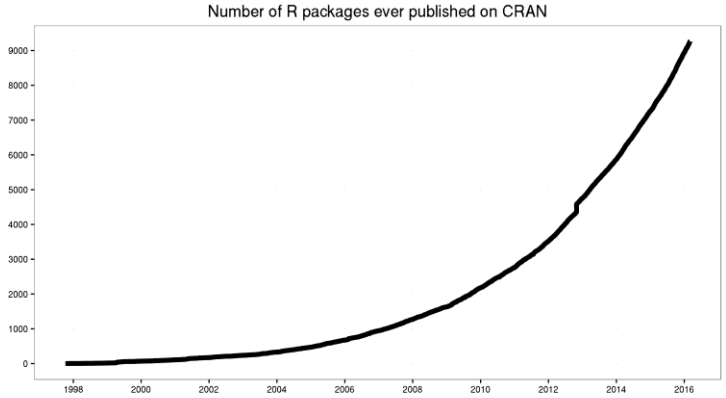
- Amazing R
- Easy R integration with SQL Server
- True power of Microsoft R product suite

R is growing

- A language platform
  - Functional language optimized for statistics and data science
  - Data visualization framework
  - Provided as Open Source
- A community
  - 3M+ statistical analysis and machine learning users
  - Taught in most university statistics programs
  - Active user groups across the world
- An ecosystem
  - CRAN: 9000+ freely available packages, test data and evaluations
  - Many applicable to big data if scaled



http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

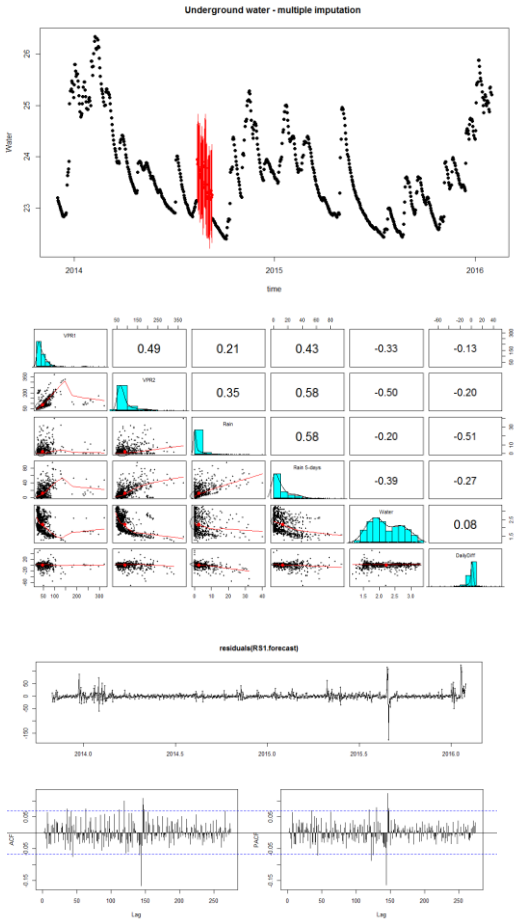


R principles

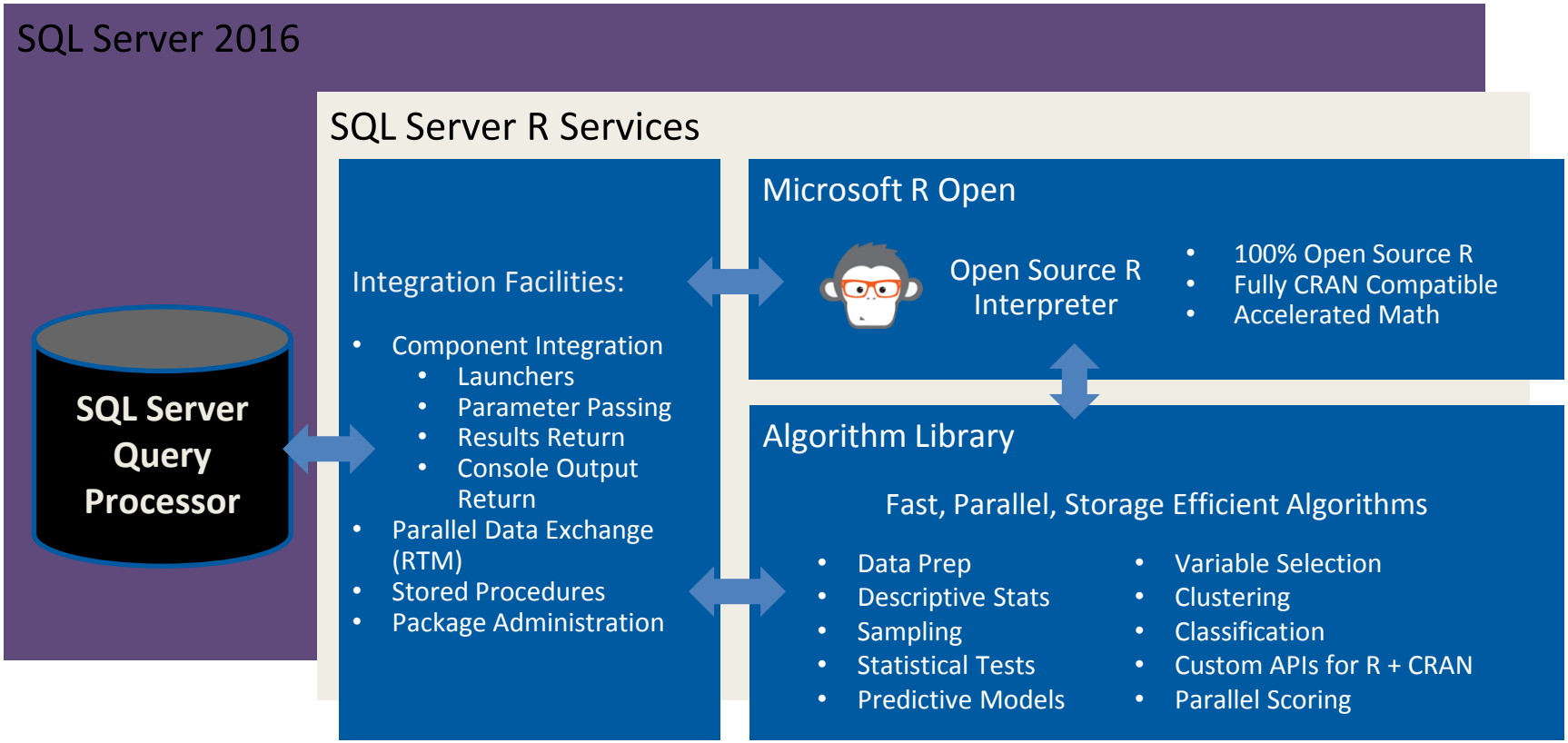
- R is an implementation of the S language
  - Developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks
  - S stands for statistics
- R has become more popular than S or S-Plus
  - It's free
  - More people are contributing to it
- R is polymorphic
  - Single function can be applied to different types of inputs, which the function processes in the appropriate way
  - If you apply the plot() function to a list of numbers, you get a simple plot
  - But if you apply it to the output of a regression analysis, you get a set of plots representing various aspects of the analysis

Demo: Time series analysis with R

- Accurate and timely forecast drives success, but it's difficult to make predictions, especially about the future
- Our goals:
  - Complete missing data
  - Measuring the effect of the rainfall and underground water level to volume pumped
  - Detect anomalies
  - Predict pumped volumes for next week



Introducing SQL Server R Services



Administering SQL Server R Services

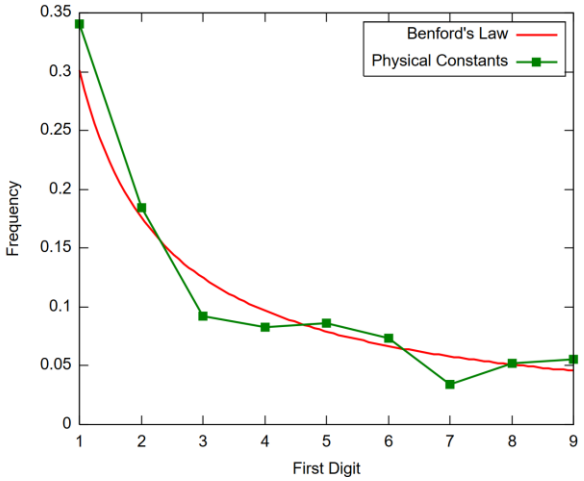
- Install and enable SQL 2016 & R Services
  - Install SQL Server 2016
  - Install either SQL Server R Services (In-Database) or the new Microsoft R Server (Standalone)
- Enable R scripts within TSQL
  - EXEC sp_configure 'external scripts enabled', 1
- Grant permission to read data
  - ALTER ROLE [db_datareader] ADD MEMBER ...
- Grant permission to run algorithms
  - ALTER ROLE [db_rrerole] ADD MEMBER ...

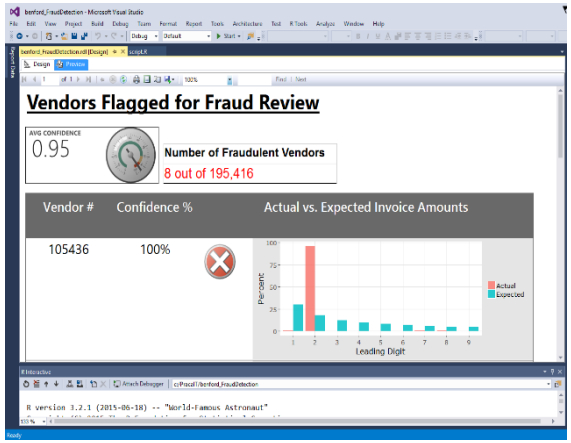
Why in-database analytics?

- The primary advantage of SQL Server R Services is data locality
- With R running in the database, you
  - Eliminate the performance hit associated with moving the data
  - Are able to encapsulate the whole operation into a stored procedure
- The gateway for this is sp_execute_external_script
  - This stored procedure allows you to pipe data from SQL Server to R using standard queries
  - An R variable, usually a data frame, can be returned back

Demo: Fraud detection with SQL Server R Services

- Fraud detection is one of the earliest industrial applications of advanced analytics
- This time we use Benford's law
- Our goals:
  - Identify fraudulent vendors based on check amounts
  - Operationalize this solution using SQL Server R Services





Revolution Analytics

- Leading commercial provider of software and services for R:
  - Revolution R OPEN
    - Enhanced with a Reproducibility Toolkit and multi-core processing
  - Revolution R Enterprise (RRE)
    - Fast, cost effective enterprise-class big data analytics platform

RevoScaleR library

- Big Data predictive analytics library included with Revolution R Enterprise
  - Now integrated with SQL Server 2016, soon with Azure HDInsight and Azure Machine Learning
- Offers enterprise-grade, terabyte-class software based on the Open Source R
  - Extremely fast statistical analysis
- http://www.rdocumentation.org/packages/RevoScaleR
- http://www.revolutionanalytics.com/sites/default/files/revoscaler-speed-scalability.pdf

RevoScaleR strengths

- Data chunking
  - RevoScaleR has its own file format (XDF)
  - Data is stored in the same binary format that is used in memory
  - Data can be accessed rapidly by row or by column
  - Blocks of contiguous rows for selected columns can be read sequentially
- Parallelism
  - Nearly all computations are automatically threaded
  - Automatic and efficient parallelization of "external memory" algorithms
  - Sophisticated algorithm for pre-analyzing models to detect data duplication

RevoScaleR compute contexts

- RevoScaleR functions can run in-Hadoop or in-Database without any functional R recoding
- Local Parallel – Linux or Windows
  - rxSetComputeContext("localpar")
- Hadoop
  - myHadoopCluster <- RxHadoopMR()
  - rxSetComputeContext(myHadoopCluster)
- SQL Server
  - mySqlServer <- RxInSqlServer()
  - rxSetComputeContext(mySqlServer)

Microsoft R product suite

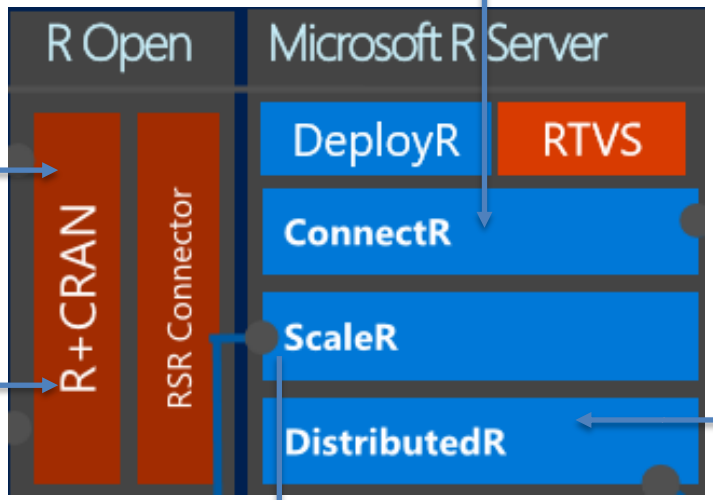
- Microsoft acquired revolution analytics in May 2015
- Microsoft R Open
  - Free and open source R distribution
    - Compatible with all R-related software
  - Enhanced and distributed by Microsoft
    - Intel MKL Library
- Microsoft R Server
  - Secure, scalable and supported distribution of R
  - With commercial components created by Microsoft

Microsoft R Open

- MRAN website
  - mran.revolutionanalytics.com
- Reproducible R toolkit
  - checkpoint, miniCRAN
- ParallelR
  - parallelise via forech loop
- Rhadoop
  - rhdfs, rhbase,ravro, rmr2, plyrmr
- AzureML
  - read/write to AzureML, publish R code as ML API

Microsoft R Server

R+ CRAN
- Open source R interpreter
- R 3.2.4
- Freely-available huge range of R algorithms
- Algorithms callable by MRO
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages



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
  - EDWs and ADWs
  - ODBC

DistributedR
- Distributed computing framework
- Delivers cross-platform portability

MRO
- 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
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Range of predictive functions
- User tools for distributing customized R algorithms across nodes
- Wide data sets supported – thousands of variables

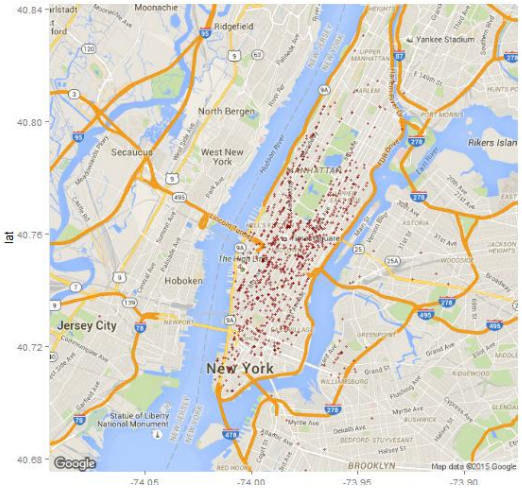
CRAN vs MRO vs MRS

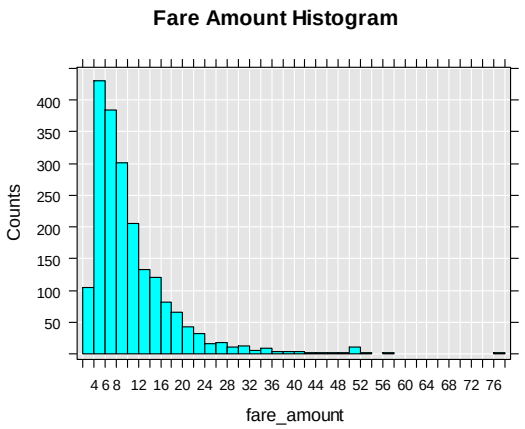
Microsoft R Server

Datasize	In-memory	In-memory	In-Memory or Disk Based
Speed of Analysis	Single threaded	Multi-threaded	Multi-threaded, paralel processing 1:N servers
Support	Community	Community	Community + Commercial
Analytic Breadth & Depth	9 000+ innovative analytic packages	9 000+ innovative analytic packages	9 000+ innovative analytic packages plus commercial parallel highspeed functions
Licence	Open Source	Open Source	Commercial license. Supported release with indemnity

Demo: Predictive analytics with SQL Server Enterprise R Services

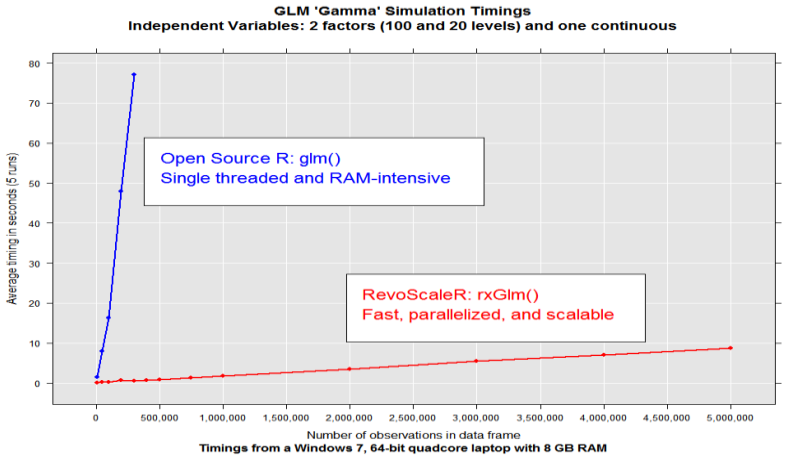
- Logit model measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function
- Five-lesson tutorial available on MSDN
- Goals:
  - Train logistic regression model to predict the probability of a driver receiving a tip for a ride
  - Evaluate the model using ROC curves
  - Deploy the prediction into SQL Server as a T-SQL stored procedure

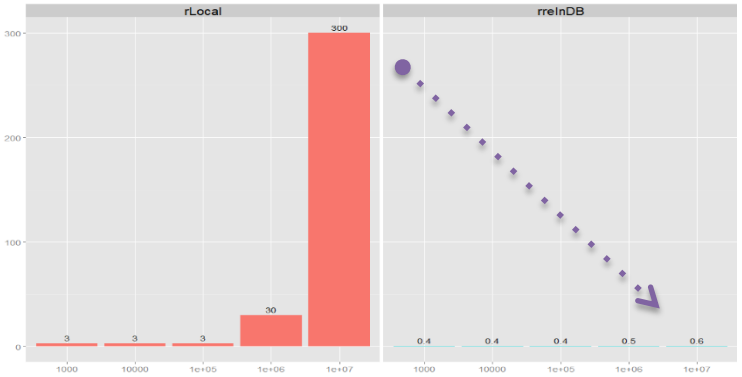




Microsoft R Server – it scales

- Scaling through parallelization
  - Scalable parallel algorithms
  - No memory limits
  - High quality R development tools
  - Web services deployment
- Achieving even greater scale by remote execution
  - Improved memory utilization
  - Remote execution in hadoop, grids, EDWs





5+ hours to 40 seconds

Summary

- R brings a lot of new possibilities ranging from data exploration to predictive modeling
- Don't miss this revolutionary opportunity
  - R is a natural fit since you already think in rows and columns
  - R has a full range of support for a wider variety of data manipulation, visualization, and machine learning techniques
  - R relies on its community for package development, and SQL pros would have access to anything new or trending in the R realm without having to wait for Microsoft to implement it
- Remeber, SQL Server R Services are much more than just ability to execute R scripts