The Cortana Intelligence Suite

Microsoft R for the SQL Professional –
Planning, Setup and Environment

Microsoft Machine Learning and Data Science Team
CortanaIntelligence.com

1. Main page: http://cortanaanalytics.com
2. To begin this module, you should have:
   1. Basic Math and Stats skills
   2. Business and Domain Awareness
   3. General Computing Background

NOTE: These workbooks contain many resources to lead you through the course, and provide a rich set of references that you can use to learn much more about these topics. If the links do not resolve properly, type the link address in manually in your web browser. If the links have changed or been removed, simply enter the title of the link in a web search engine to find the new location or a corollary reference.

- Log in to the site:

  http://cisw-foundations.azurewebsites.net/

- Select the "The Cortana Intelligence Suite – Foundations (Course Materials)" Link

- Click the "Course Materials" button, and then "Sign in…"

- Open "Edit in Browser", Find the "R For T-SQL" tab, Enter your primary role at your company on the sign-in sheet.

Welcome!
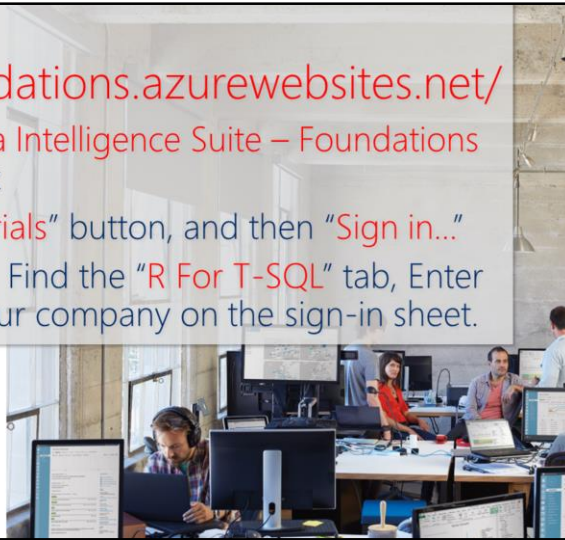
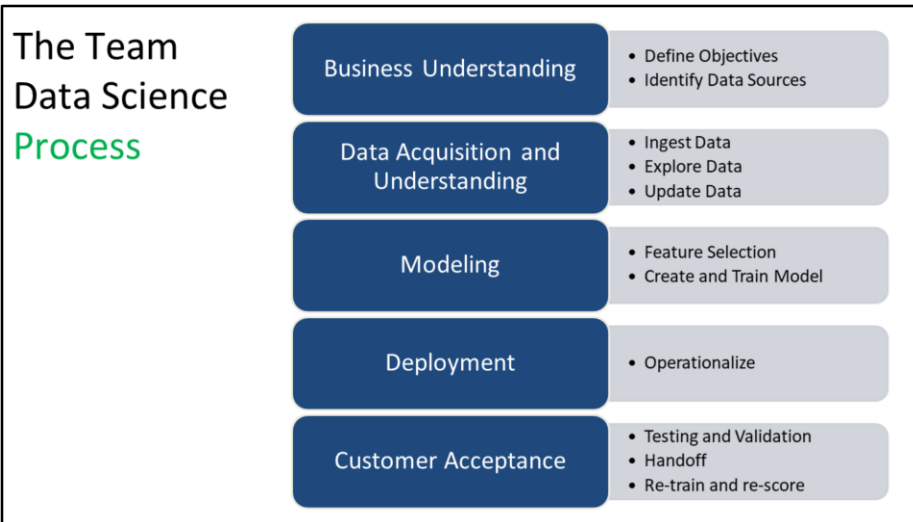1. At the end of this Module, you will:
    1. Understand the R Language and where it is used
    2. Understand the Microsoft R Platform and its capabilities
    3. Set up and use the server and various client tools for a R environment

The Data Science Process and Platform

1. This process largely follows the CRISP-DM model: http://www.sv-europe.com/crisp-dm-methodology/
2. It also references the Cortana Intelligence process: https://azure.microsoft.com/en-us/documentation/articles/data-science-process-overview/
3. A complete process diagram is here: https://azure.microsoft.com/en-us/documentation/learning-paths/cortana-analytics-process/
4. Some walkthrough's of the various services: https://azure.microsoft.com/en-us/documentation/articles/data-science-process-walkthroughs/
5. An integrated process and toolset allows for a more close-to-intent deployment
6. Iterations are required to close in on the solution – but are harder to management and monitor

Data Science Blog: https://buckwoody.wordpress.com/

1. Platform and Storage: Microsoft Azure – http://microsoftazure.com Storage: https://azure.microsoft.com/en-us/documentation/services/storage/ (Host It)
2. Azure Data Catalog: http://azure.microsoft.com/en-us/services/data-catalog (Doc It)
3. Azure Data Factory: http://azure.microsoft.com/en-us/services/data-factory/ (Move It)
4. Azure Event Hubs: http://azure.microsoft.com/en-us/services/event-hubs/ (Bring It)
5. Azure Data Lake: http://azure.microsoft.com/en-us/campaigns/data-lake/ (Store It)
6. Azure DocumentDB: https://azure.microsoft.com/en-us/services/documentdb/ , Azure SQL Data Warehouse: http://azure.microsoft.com/en-us/services/sql-data-warehouse/ (Relate It)
7. Azure Machine Learning: http://azure.microsoft.com/en-us/services/machine-learning/ (Learn It)
8. Azure HDInsight: http://azure.microsoft.com/en-us/services/hdinsight/ (Scale It)
9. Azure Stream Analytics: http://azure.microsoft.com/en-us/services/stream-analytics/ (Stream It)
10. Power BI: https://powerbi.microsoft.com/ (See It)
11. Cortana: http://blogs.windows.com/buildingapps/2014/09/23/cortana-integration-and-speech-recognition-new-code-samples/ and https://blogs.windows.com/buildingapps/2015/08/25/using-cortana-to-interact-with-your-customers-10-by-10/ and https://developer.microsoft.com/en-us/Cortana (Say It)
12. Cognitive Services: https://www.microsoft.com/cognitive-services
13. Bot Framework: https://dev.botframework.com/
14. All of the components within the suite: https://www.microsoft.com/en-us/server-cloud/cortana-intelligence-suite/what-is-cortana-intelligence.aspx
15. What can I do with it? https://gallery.cortanaintelligence.com/
16. Getting Started Quickly: https://caqs.azure.net/#gallery

The R Ecostructure

1. Video Introduction to R: https://mran.revolutionanalytics.com/documents/what-is-r/

1. One-Page R: https://togaware.com/onepager/
2. R on Youtube: https://www.youtube.com/user/thelearnr
3. R Links: http://www.datasciencecentral.com/m/discussion?id=6448529%3ATopic%3A280135
4. R resources: https://msdn.microsoft.com/en-us/microsoft-r/microsoft-r-more-resources

SQL and R Contrasted

1011
101

SQL

1. Client/Server
2. Database Objects
3. DML, DDL
4. DCL
5. Declarative Code

R

1. Interactive Environment
2. Data Structures
3. Functions
4. Libraries (Packages)
5. Functional Code Flow

1. Learn SQL: http://www.w3schools.com/SQl/default.asp
2. Try R, with a great interface. http://tryr.codeschool.com/levels/1/challenges/22
3. R and Statistics Intro: https://www.youtube.com/watch?v=xb5P5xdcr2U&feature=youtu.be&a
4. R Online: http://www.tutorialspoint.com/r_terminal_online.php
5. Using R to explore data: http://www.analyticsvidhya.com/blog/2015/10/cheatsheet-11-steps-data-exploration-with-codes/
6. Quick R Intro: http://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A112754
7. Creating a recommender engine in R: http://www.analyticbridge.com/profiles/blogs/build-basic-recommendation-engine-using-r
8. Visualizations cheat-sheet in R: http://www.datasciencecentral.com/forum/topics/cheat-sheet-data-visualization-with-r?groupUrl=tutorials
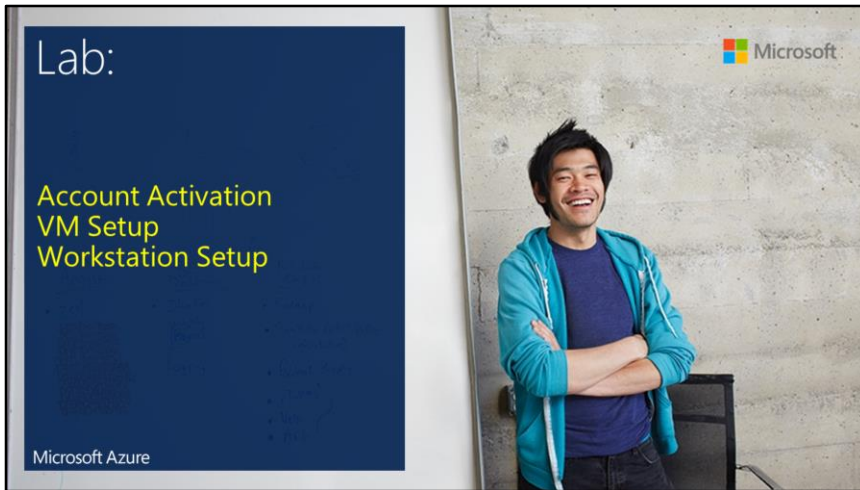
1. If you do not have a Microsoft Azure account, go here: https://azure.microsoft.com/en-us/free/ (You will need a credit card, but you will not be charged)
    1. Log in to the Azure Portal: https://ms.portal.azure.com
    2. Create a new Windows Data Science Virtual Machine (2 Processors, 7GB RAM, HDD): https://azure.microsoft.com/en-us/documentation/articles/machine-learning-data-science-vm-do-ten-things/
2. *Optional, if using your local machine:*
    1. Install SQL Server 2016 and ensure you select R Services – see this link: https://www.microsoft.com/en-us/cloud-platform/sql-server-editions-developers
    2. Install Visual Studio Community Edition 2015: https://www.microsoft.com/en-us/download/details.aspx?id=48146
    3. Install SQL Server Data Tools: https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt
    4. Install R Tools for Visual Studio: https://microsoft.github.io/RTVS-docs/

The Microsoft R Platform

1. Primary Microsoft R Site: https://msdn.microsoft.com/en-us/microsoft-r/index

Microsoft R Products

Microsoft R Open

- Free and open source R distribution
- Enhanced and distributed by Revolution Analytics

SQL Server R Services

- Built in Advanced Analytics and Stand Alone Server Capability
- Leverages the Benefits of SQL 2016 Enterprise Edition

Microsoft R Server

- Microsoft R Server for Redhat Linux
- Microsoft R Server for SUSE Linux
- Microsoft R Server for Teradata DB
- Microsoft R Server for Hadoop on Redhat

1. Channel 9 videos on Microsoft R:
   https://channel9.msdn.com/Search?term=Microsoft%20R#lang-en=en&ch9Search

Microsoft R Open

- Enhanced Open Source R distribution
  - Based on the latest Open Source R
  - Built, tested and distributed by Microsoft
  - Enhanced by Intel MKL Library to speed up linear algebra functions
- Compatible with all R-related software
  - CRAN packages, RStudio, third-party R integrations, …
- Revolutions Open-Source R packages
  - Reproducible R Toolkit – checkpoint
- MRAN website mran.revolutionanalytics.com
  - Enhanced documentation and learning resources
  - Discover 7500 free add-on R packages
- Open source (GPLv2 license) - 100% free to download, use and share

1. Quick Video on R Client: https://channel9.msdn.com/blogs/MicrosoftR/Microsoft-Introduces-new-free-Microsoft-R-Client

CRAN R compared to Microsoft R Open

Performance comparison

- R-3.1.1 (1-core)
- RRO (1-core)
- RRO (4-core)

Urbanek benchmark execution time

- Matrix calculation – up to 27x faster
- Matrix functions – up to 16x faster
- Programation – 0x faster
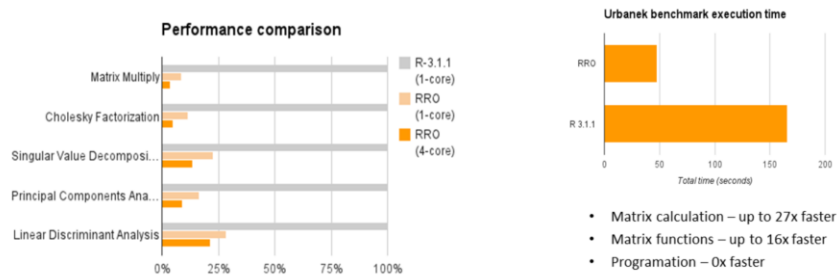
- More efficient and multi-threaded math computation.
- Benefits math intensive processing.
- No benefit to program logic and data transform

1. Overview: https://channel9.msdn.com/Series/Microsoft-R-Server-Series/Introduction-to-Microsoft-R-Server-Session-1--Overview

1. Book and Series: http://dacrook.com/introduction-to-microsoft-r-open/
2. Microsoft R Client: https://msdn.microsoft.com/en-us/microsoft-r/index#mrc

Microsoft R Components
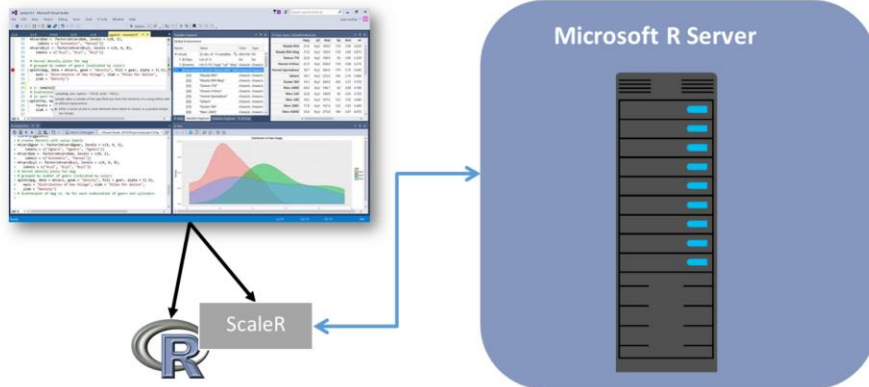
CRAN R
- Microsoft R Open
- Microsoft R Client
- Microsoft R Server
- HDInsight SparkR / SQL Server R Services

- R in Azure Machine Learning

1. Supported Platforms for Microsoft R Server: https://msdn.microsoft.com/en-us/microsoft-r/rserver-install-supported-platforms
2. Book and Series: http://dacrook.com/introduction-to-microsoft-r-open/
3. Microsoft R Client: https://msdn.microsoft.com/en-us/microsoft-r/index#mrc
4. Microsoft R Server: https://msdn.microsoft.com/en-us/microsoft-r/index#mrs
5. SQL Server R Services: https://msdn.microsoft.com/en-us/microsoft-r/index#sqlr
6. HDInsight SparkR: https://azure.microsoft.com/en-gb/services/hdinsight/apache-spark/

Microsoft R Server

Microsoft R Server is a broadly deployable enterprise-class analytics platform based on R that is supported, scalable and secure. Supporting a variety of big data statistics, predictive modeling and machine learning capabilities, R Server supports the full range of analytics – exploration, analysis, visualization and modeling

High-performance open source R plus:

- Data source connectivity to big-data objects
- Big-data advanced analytics
- Multi-platform environment support
- Inpredictive modeling
- Development and production environment support
  - IDE for data scientist developers
  - Secure, Scalable R Deployment

R Open	R Server
R+CRAN / RSR Connector	ConnectR
	ScaleR
	DistributedR
Open Source Components	Licensed Components

1.	Microsoft R Server: https://msdn.microsoft.com/en-us/microsoft-r/index#mrs
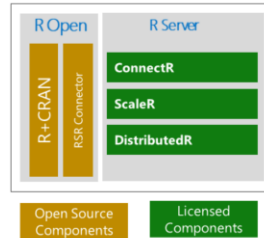
The Microsoft R Server Platform

R+CRAN
- Open source R interpreter
  - R 3.2.2
- Freely-available huge range of R algorithms
- Algorithms callable by MSR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

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

R Open | Microsoft R Server

R+CRAN | RSR Connector

ConnectR
ScaleR
DistributedR

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

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

DistributedR
- Distributed computing framework
- Delivers cross-platform portability

1. Installing on Linux: https://channel9.msdn.com/Series/Microsoft-R-Server/Microsoft-R-Server-Installation-Linux

CRAN, MRO, MRS Comparison

	![R]	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**
License	Open Source	Open Source	**Commercial license. Supported release with indemnity**

1. Technology Overview: https://channel9.msdn.com/Series/Microsoft-R-Server/Technology-Overview-for-Microsoft-R-Server-2016

CRAN R	MRO	MRS

Adds multithreaded computation for linear algebra

Adds functions and infrastructure for scalable computation over local and distributed platforms

Makes it easy to share R code and replicate results using specific R package versions.

MRS Packages
- doRSR
- RevolOQ
- RevoMods
- RevoPemaR
- RevoRpeConnector
- RevoRsrConnector
- RevoScaleR
- RevoTreeView
- RevoUtils

MRO Packages
- checkpoint (also on CRAN)
- RevoUtilsMath (for MKL)

MRO Packages
- checkpoint (also on CRAN)
- RevoUtilsMath (for MKL)

Standard Packages
- base
- recommended

Standard Packages
- base
- recommended

Standard Packages
- base
- recommended

BLAS / LAPACK - Standard | BLAS / LAPACK – Intel MKL | BLAS / LAPACK – Intel MKL

R Kernel | R Kernel | R Kernel

1. Getting Started: https://msdn.microsoft.com/en-us/microsoft-r/?f=255&MSPPError=-2147217396

1. Function Breakdown: https://msdn.microsoft.com/en-us/microsoft-r/scaler/scaler

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
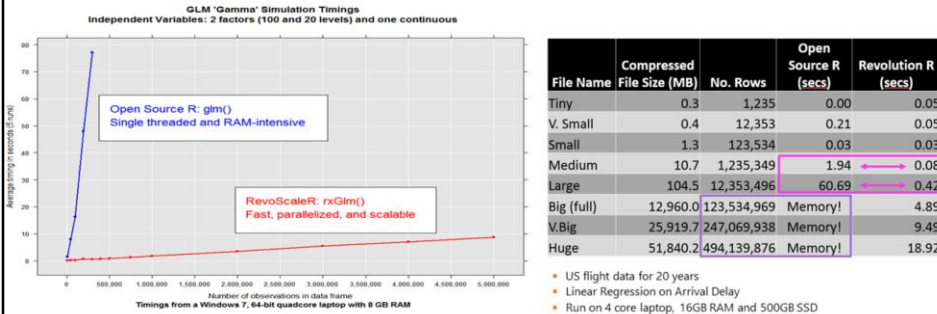- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms

1. SQL Server Implementation of ScaleR Functions: https://msdn.microsoft.com/en-us/library/mt652103.aspx
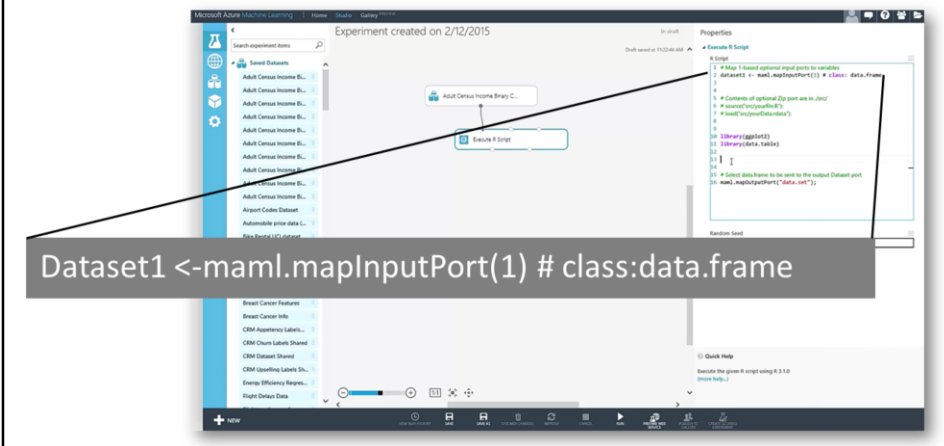
ScaleR - Performance comparison

Microsoft R Server has no data size limits in relation to size of available RAM. When open source R operates on data sets that exceed RAM it will fail. In contrast Microsoft R Server scales linearly well beyond RAM limits and parallel algorithms are much faster.
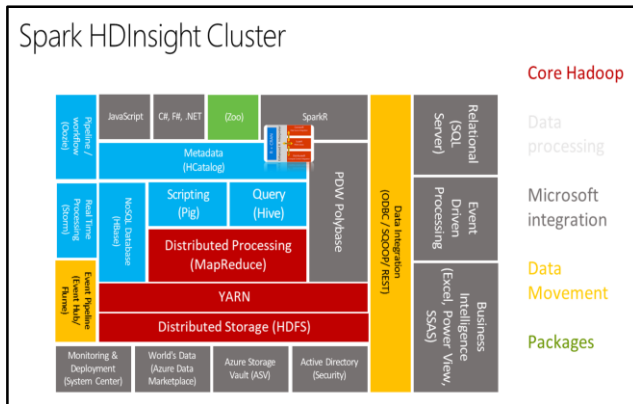
GLM 'Gamma' Simulation Timings
Independent Variables: 2 factors (100 and 20 levels) and one continuous

Open Source R: glm()
Single threaded and RAM-intensive

RevoScaleR: rxGlm()
Fast, parallelized, and scalable

Number of observations in data frame
Timings from a Windows 7, 64-bit quadcore laptop with 8 GB 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

1. ScaleR Functions for Working with SQL Server Data: https://msdn.microsoft.com/en-us/library/mt732681.aspx
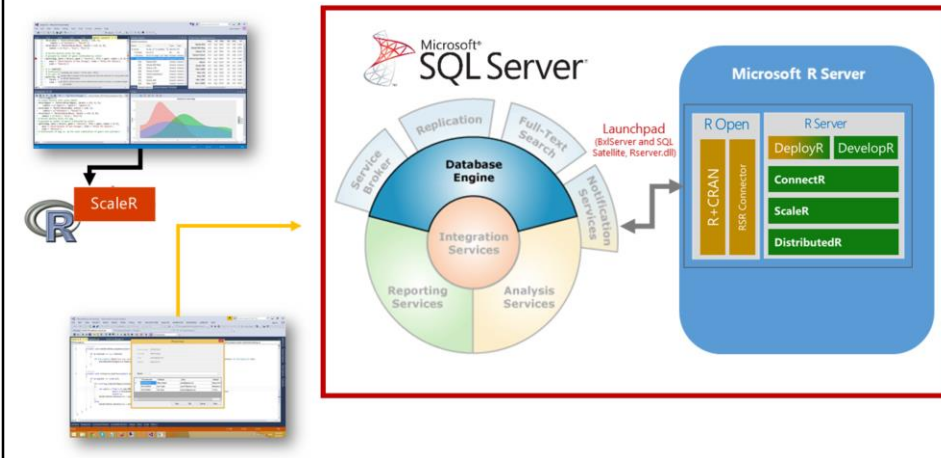
1. Primary reference: https://msdn.microsoft.com/en-us/library/dn905952.aspx
2. Using R in Azure Machine Learning: https://azure.microsoft.com/en-us/documentation/articles/machine-learning-r-quickstart/
3. Overview Video: https://channel9.msdn.com/Blogs/Windows-Azure/R-in-Azure-ML-Studio
4. R Packages supported: https://msdn.microsoft.com/en-us/library/mt741980.aspx

1. Full training example for the local HDP Instance: http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/

2. More detail on the Hadoop Components: http://www.datasciencecentral.com/profiles/blogs/hadoop-herd-when-to-use-what

1. Primary Documentation and training: https://msdn.microsoft.com/en-us/library/mt604845.aspx

R Client Options

1. The Microsoft R Client: https://msdn.microsoft.com/en-us/microsoft-r/install-r-client-windows
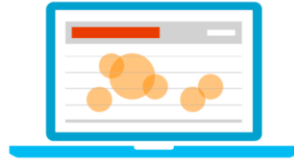
Microsoft Data Science

Microsoft R Development Tools

- Command-Line
- RStudio
- R Tools for Visual Studio (RTVS)
- SQL Server tools

1. Installing Microsoft R Client on Windows: https://msdn.microsoft.com/en-us/microsoft-r/install-r-client-windows
2. Files located at: C:\Program Files\Microsoft\R Client\R_SERVER\bin

The R Environment

- Profiles
- Version Control
- Package Locations
- Workspaces

Note: "Environment" has a specific meaning in R – we are using the general term environment to mean the operating surroundings for your computer running R - http://adv-r.had.co.nz/Environments.html#env-basics

1. Information on the R Profiles: https://www.r-bloggers.com/fun-with-rprofile-and-customizing-r-startup/

1. Open Visual Studio, and read through the walkthrough of the RTVS tools for SQL Server and R: https://microsoft.github.io/RTVS-docs/sqlserver.html
2. Download the **Resources** zip file as described by your instructor and extract all the files to a local directory on your system.
3. Open the file **R for SQL Professionals Lab (Student).R** and complete all exercises in **#1.0 Planning, setup and environment**. Look for three # symbols for the tasks you should complete. Work through all exercises and stop at **#1.4 Package Management**.
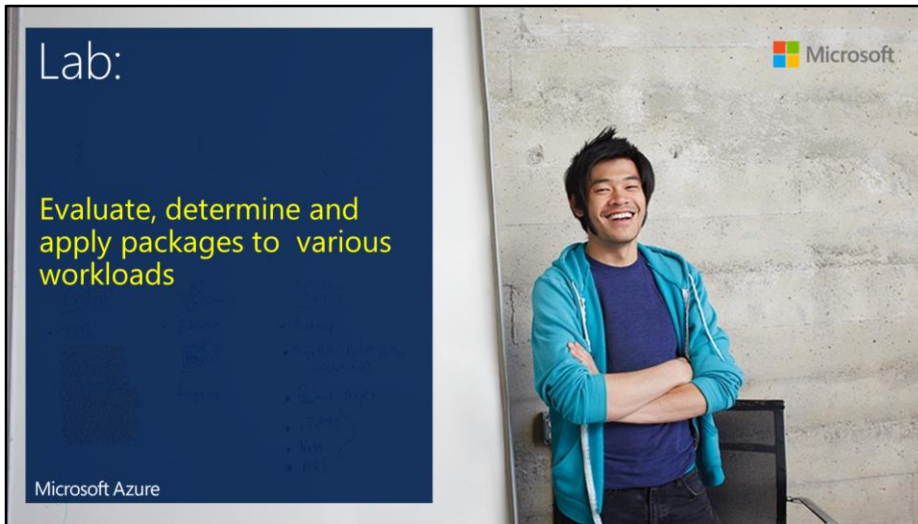
R Package Management

1. Video Introduction to R: https://mran.revolutionanalytics.com/documents/what-is-r/

Packages

- Use
- Adding
- Querying
- Considerations

1. Packages: http://www.dummies.com/programming/r/how-to-install-load-and-unload-packages-in-r/ and https://cran.r-project.org/doc/manuals/R-admin.html#Add_002don-packages
2. Book on Creating your own Packages: http://r-pkgs.had.co.nz/
3. A useful set of packages: https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages
4. R Packages supported by Azure Machine Learning: https://msdn.microsoft.com/en-us/library/mt741980.aspx
5. R Package Management for SQL Server R Services: https://msdn.microsoft.com/en-us/library/mt790486.aspx
6. Scaling Packages: https://msdn.microsoft.com/en-US/library/mt637368.aspx

1. Open the file R for SQL Professionals Lab (Student).R and complete #1.4 Package Management.
2. Check to see if the following packages are installed:
   1. dplyr
   2. tidyr
   3. lubridate
   4. ggplot2
   5. xtable
   6. maps
   7. zoo
   8. knitr
3. If any of these are not installed, install them
4. Open the help for each of these libraries and work through one sample

Microsoft Data Science

Questions?

More resources:
1. https://msdn.microsoft.com/en-us/microsoft-r/microsoft-r-more-resources
2. Revolutions Blog
3. Blog: Joseph Sirosh, "Making R the Enterprise Standard..."
4. Getting Started with Microsoft R
5. Diving In.. Data Analysis in Microsoft R
6. R Server Technology – Video
7. R Tools for Visual Studio Sneak Peek
8. R Tools for Visual Studio Overview
9. SQL R Services Overview – Youtube
10. SQL R Services Feature Overview - Youtube
11. SQL R Services Overview at Build
12. SQL R Services Tutorial