R developer's guide to Azure

03/19/2019 • 9 minutes to read • Contributors

In this article

Many data scientists dealing with ever-increasing volumes of data are looking for ways to harness the power of cloud computing for their analyses. This article provides an overview of the various ways that data scientists can leverage their existing skills with the R programming language in Azure.

Microsoft has fully embraced the R programming language as a first-class tool for data scientists. By providing many different options for R developers to run their code in Azure, the company is enabling data scientists to extend their data science workloads into the cloud when tackling large-scale projects.

Let's examine the various options and the most compelling scenarios for each one.

Azure services with R language support

This article covers the following Azure services that support the R language:

Service	Description
Data Science Virtual Machine	a customized VM to use as a data science workstation or as a custom compute target
ML Services on HDInsight	cluster-based system for running R analyses on large datasets across many nodes
Azure Databricks	collaborative Spark environment that supports R and other languages
Azure Machine Learning Studio	run custom R scripts in Azure's machine learning experiments
Azure Batch	offers a variety options for economically running R code across many nodes in a cluster
Azure Notebooks	a no-cost cloud-based version of Jupyter notebooks
Azure SQL Database	run R scripts inside of the SQL Server database engine

Data Science Virtual Machine

The Data Science Virtual Machine (DSVM) is a customized VM image on Microsoft's Azure cloud platform built specifically for doing data science. It has many popular data science tools, including:

- Microsoft R Open

- Microsoft Machine Learning Server
- RStudio Desktop
- RStudio Server

The DSVM can be provisioned with either Windows or Linux as the operating system. You can use the DSVM in two different ways: as an interactive workstation or as a compute platform for a custom cluster.

As a workstation

If you want to get started with R in the cloud quickly and easily, this is your best bet. The environment will be familiar to anyone who has worked with R on a local workstation. However, instead of using local resources, the R environment runs on a VM in the cloud. If your data is already stored in Azure, this has the added benefit of allowing your R scripts to run "closer to the data." Instead of transferring the data across the Internet, the data can be accessed over Azure's internal network, which provides much faster access times.

The DSVM can be particularly useful to small teams of R developers. Instead of investing in powerful workstations for each developer and requiring team members to synchronize on which versions of the various software packages they will use, each developer can spin up an instance of the DSVM whenever needed.

As a compute platform

In addition to being used as a workstation, the DSVM is also used as an elastically scalable compute platform for R projects. Using the AzureDSVM R package, you can programmatically control the creation and deletion of DSVM instances. You can form the instances into a cluster and deploy a distributed analysis to be performed in the cloud. This entire process can be controlled by R code running on your local workstation.

To learn more about the DSVM, see Introduction to Azure Data Science Virtual Machine for Linux and Windows.

ML Services on HDInsight

Microsoft ML Services provide data scientists, statisticians, and R programmers with on-demand access to scalable, distributed methods of analytics on HDInsight. This solution provides the latest capabilities for R-based analytics on datasets of virtually any size, loaded to either Azure Blob or Data Lake storage.

This is an enterprise-grade solution that allows you to scale your R code across a cluster. By leveraging functions in Microsoft's RevoScaleR package, your R scripts on HDInsight can run data processing functions in parallel across many nodes in a cluster. This allows R to crunch data on a much larger scale than is possible with single-threaded R running on a workstation.

This ability to scale makes ML Services on HDInsight a great option for R developers with massive data sets. It provides a flexible and scalable platform for running your R scripts in the cloud.

For a walk-through on creating an ML Services cluster, see Get started with ML Services on Azure HDInsight.

Azure Databricks

Azure Databricks is an Apache Spark-based analytics platform optimized for the Microsoft Azure cloud services platform. Designed with the founders of Apache Spark, Databricks is integrated with Azure to provide one-click setup, streamlined workflows, and an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

The collaboration in Databricks is enabled by the platform's notebook system. Users can create, share, and edit notebooks with other users of the systems. These notebooks allow users to write code that executes against Spark

clusters managed in the Databricks environment. These notebooks fully support R and give users access to Spark through both the `SparkR` and `sparklyr` packages.

Since Databricks is built on Spark and has a strong focus on collaboration, the platform is often used by teams of data scientists that work together on complex analyses of large data sets. Because the notebooks in Databricks support other languages in addition to R, it is especially useful for teams where analysts use different languages for their primary work.

The article What is Azure Databricks? can provide more details about the platform and help you get started.

Azure Machine Learning Studio

Azure Machine Learning Studio is a collaborative, drag-and-drop tool you can use to build, test, and deploy predictive analytics solutions in the cloud. It enables emerging data scientists to create and deploy machine learning models without the need to write much code.

Azure Machine Learning Studio supports both R and Python. You can use R with Azure Machine Learning Studio in two ways.

Custom R scripts in your experiments

First, you can extend the data manipulation and machine learning capabilities of ML Studio by writing custom R scripts. Although ML Studio includes a wide variety of modules for preparing and analyzing data, it cannot match the capabilities of a mature language like R. Therefore, the service was designed to allow you to introduce your own custom R scripts in cases where the provided modules do not meet your needs.

To leverage this capability, drag and drop an "Execute R Script" module into your experiment. Then use the code editor in the "Properties" pane to write a new R script or paste an existing script. Within the script, you can reference external R packages. You can use the script to manipulate data or to train complex ML models that are not part of the standard Azure Machine Learning Studio model library.

For a thorough introduction on using R within ML Studio experiments, check out Getting started with the R programming language in Azure Machine Learning Studio.

Create, manage, and deploy experiments from your local R environment

The other way that you can use R with Azure Machine Learning Studio is to use the AzureML package to monitor and control the experimentation process with the R programming environment. This package, which is maintained by Microsoft, allows you to upload and download datasets to and from Azure Machine Learning Studio, to interrogate experiments, to publish R functions as web services, and to run R data through existing web services and retrieve the output.

This package makes it much easier to use Azure Machine Learning Studio as a scalable deployment platform for your R code. Instead of clicking and dragging in the UI, you can automate the entire deployment process using tools you already know.

Azure Batch

For large-scale R jobs, you can use Azure Batch. This service provides cloud-scale job scheduling and compute management so you can scale your R workload across tens, hundreds, or thousands of virtual machines. Since it is a generalized computing platform, there a few options for running R jobs on Azure Batch.

One option is to use Microsoft's doAzureParallel package. This R package is a parallel backend for the `foreach` package. It allows each iteration of the `foreach` loop to run in parallel on a node within the Azure Batch cluster. For an

introduction to the package, see the blog post doAzureParallel: Take advantage of Azure's flexible compute directly from your R session.

Another option for running an R script in Azure Batch is to bundle your code with "RScript.exe" as a Batch App in the Azure portal. For a detailed walk-through, consult R Workloads on Azure Batch.

A third option is to use the Azure Distributed Data Engineering Toolkit (AZTK), which allows you to provision on-demand Spark clusters using Docker containers in Azure Batch. This provides an economical way to run Spark jobs in Azure. By using SparklyR with AZTK, your R scripts can be scaled out in the cloud easily and economically.

Azure Notebooks

Azure Notebooks is a low-cost, low-friction method for R developers who prefer working with notebooks to bring their code to Azure. It is a free service for anyone to develop and run code in their browser using Jupyter, which is an open-source project that enables combing markdown prose, executable code, and graphics onto a single canvas.

The free service tier of Azure Notebooks is a viable option for small-scale projects, as it limits each notebook's process to 4GB of memory and 1GB data sets. If you need compute and data power beyond these limitations, however, you can run notebooks in a Data Science Virtual Machine instance. For more information, see Manage and configure Azure Notebooks projects - Compute tier.

Azure SQL Database

Azure SQL Database is Microsoft's intelligent, fully managed relational cloud database service. It allows you to use the full power of SQL Server without any hassle of setting up the infrastructure. This includes Machine Learning Services in SQL Server, which is one of the more recent additions to SQL.

This feature offers an embedded, predictive analytics and data science engine that can execute R code within a SQL Server database as stored procedures, as T-SQL scripts containing R statements, or as R code containing T-SQL. Instead of extracting data from the database and loading it into the R environment, you load your R code directly into the database and let it run right alongside the data.

While Machine Learning Services has been part of on-premises SQL Server since 2016, it is relatively new to Azure SQL Database. It is currently in limited preview but will continue to evolve.

Next steps

- Running your R code on Azure with mrsdeploy
- Machine Learning Server in the Cloud
- Additional Resources for Machine Learning Server and Microsoft R
- R on Azure - an overview of packages, tools, and case studies for using R with Azure

---