# Course Syllabi Template

**(Title)**

**Scalable Data Science with Microsoft R Server + Spark and HDInsight**

**(*Specify a 2-line description of the course – example given below)***

**In this course, you’ll gain hands-on experience with Microsoft R and HDInsight Spark for scalable data science and machine learning. You will learn about the premium offering of the HDInsight platform and how to leverage Microsoft R Server as an application on top of HDInsight Spark to perform data analysis and machine learning at scale. This is a three-day course and will teach you about Microsoft R and Spark from the ground up.**

**About the Course**

*(Specify a detailed Abstract for this course, it should include details of what will be taught and by the end of the course what the student will learn)*

In this course, you’ll gain hands-on experience with Microsoft R and HDInsight Spark for scalable data science and machine learning. You will learn about the fundamentals of functional programming, parallel external memory algorithms, Spark on HDInsight, and distributed systems. This course emphasizes robust programming principles, so that you can write programs that are portable, platform invariant, and scalable. Through labs and instructor led deep dives, you will learn how to use R Server on Spark with the HDInsight platform to perform data analysis and machine learning at scale. By the end of the course, you will have developed applications that are scalable and portable, and know how to configure your Spark clusters to maximize your application's performance.

**Prerequisites**

*(Specify the pre-requisites for this course)*

There are a few things you will need in order to properly follow the course materials:

* A subscription to Microsoft Azure (this may be provided through your company or as part of your invitation – you \*must\* have this enabled prior to class. You will be using Azure throughout the course, for all labs, work, and exercises. You can use your MSDN subscription (https://azure.microsoft.com/en-us/pricing/member-offers/msdn-benefits/), your employer may provide Azure resources to you, or you may receive instructions in your class invitation, and have at least $50 to spend for the course. Understanding of R - ability to write functions, an ability to train models, etc. Putty, Cygwin, or some bash emulator (some Linux experience to go with it would be useful) It’s also a good idea to have a general level of predictive and classification modeling, and a basic understanding of Statistics and Machine Learning, i.e., cross-validation, ensemble models, model metrics, etc.

**Agenda**

*(Add a detailed agenda of what topics are covered during training, an example is given below)*

What will you learn Functional-Object Based Computing with R - Overview of the R Project and CRAN - Exploring the Microsoft R Data Stack - Functional Programming for Data Manipulation with the dplyr package - Understanding dplyr's symantics and the magrittr pipe - Data Visualization and Exploratory Data Analysis - Using the broom package for Modeling and Summarization Breaking the Memory Barrier with RevoScaleR - Overview of the Microsoft R Data Ecosystem - Modeling and Scoring with High-Performance ScaleR Algorithms - Data Manipulation with the dplyrXdf Package - Summarizing Data with RevoScaleR - Performance Considerations with RevoScaleR - Parallel Computing and Distributed Computing with Microsoft R Server - Deploying R and ScaleR algorithms to Azure with the AzureML package - Overview of the Apache Spark Project - Ingesting Data into Azure Blob Storage - Creating Spark DataFrames and Spark Contexts - Manipulating HDFS data with the sparklyr package - Creating Distributed eXternal DataFrames in HDFS - Preparing Data for Modeling with Microsoft R Server - Training Statistical Models with Microsoft R Server and the Spark Compute Context - Scoring and Deploying Models - Performance Considerations on Hadoop Skills taught - Understand what is Spark and why it's a more effective solution for iterative machine learning jobs than Hadoop MapReduce. - Understand functional programming and lazy evaluation. - Provision and deploy HDInsight Spark Clusters and install R Server as an application. - Understand the basics of administration and management of packages and applications on premium HDInsight Spark clusters. - Develop functions that are robust to different data structures and execution environments. - Use Spark and it's R APIs for exploratory data analysis. - Train and tune statistical machine learning models with Microsoft R Server's RxSpark compute context. - Deploy trained R models as an Azure ML web service.

**Author**

*(Specify the name of the author)*

**Format**

*(Specify if the training is Virtual, Classroom, MOOC, Tutorial, Video)*

**Duration**

*(Specify the number of days this training would take to complete)*

**Course Topics**

*(Specify the product technologies covered for this training)*

**Intended Audience**

***(****Provide a brief description who this training is intended for, example given here)*

General, Data Scientists, Architects  
This course is meant for data scientists, data analysts, and experienced data architects who have programming experience with R and want to use it with Hadoop and Spark. Participants are expected to know the following about R: it's data structures, creating and using functions, and a little bit about functional programming. Participants are expected to know the basics of Azure Data Storage, and will need to have an Azure subscription with at least $50 to spend to complete this course. Some understanding of Hadoop and Spark is recommended, but not required.

**Course Level**

*(Specify if the course level is a Beginner or Intermediate or Advanced or Certificate)*