# Instructions for preparing Linux HDI Spark clusters running MRS and performing analytical tasks described in the KDD 2016 tutorial

**John-Mark Agosta**, **Debraj GuhaThakurta**, **Robert Horton**, **Mario Inchiosa, Srini Kumar**,

**Vanja Paunic**, **Hang Zhang,** **Mengyue Zhao**

Microsoft Corporation

## KDD 2016 tutorial: Scalable R on Spark

In August 2016, we presented a tutorial titled: (links to [KDD tutorial page](http://www.kdd.org/kdd2016/tutorials/view/scalable-r-on-spark); [ACM digital library](http://dl.acm.org/citation.cfm?id=2945398&dl=ACM&coll=DL&CFID=839147418&CFTOKEN=52861733), and [public GitHub repository for tutorial materials](https://github.com/Azure/Azure-MachineLearning-DataScience/tree/master/Misc/KDDCup2016)). This document describes the scripts that are necessary for creating, preparing and deleting HDInsight Spark clusters on which the tutorial exercises can be executed.

## Overview of what is provided in this document

Here we describe the scrips for preparing Spark HDInsight clusters, as well as the codes that were used for the exercises during the KDD 2016 tutorial. For preparing the clusters we provide [Windows PowerShell](https://technet.microsoft.com/en-us/library/bb978526.aspx) scripts with [Azure commandlets](https://msdn.microsoft.com/en-us/library/azure/jj554330.aspx), a set of Windows PowerShell modules that help you to automate your Microsoft Azure tasks.

Here we provide details of the following steps:

1. Provisioning Azure HDInsight Spark clusters with MRS and the necessary blob storage accounts for the clusters
2. Installing necessary software and R packages using script actions
3. Testing clusters for validity
4. Copying necessary files needed to perform the exercises described in the tutorial
5. Execution of the R codes provided for the tutorial (other R codes should also run on the same clusters)
6. Deleting clusters and associated storage accounts

For steps 1, 2, and 6, we provide scripts that will run in parallel, thereby reducing the total cluster preparation and deletion times.

All the scripts and code are available for download from a [public GitHub repository](https://github.com/Azure/Azure-MachineLearning-DataScience/tree/master/Misc/KDDCup2016). Scripts for creating, preparing, and deleting clusters are in the folder “Scripts”. R code for exercises are in the folder “Code”. Slides presented during the tutorial are in the folder “Slides”.

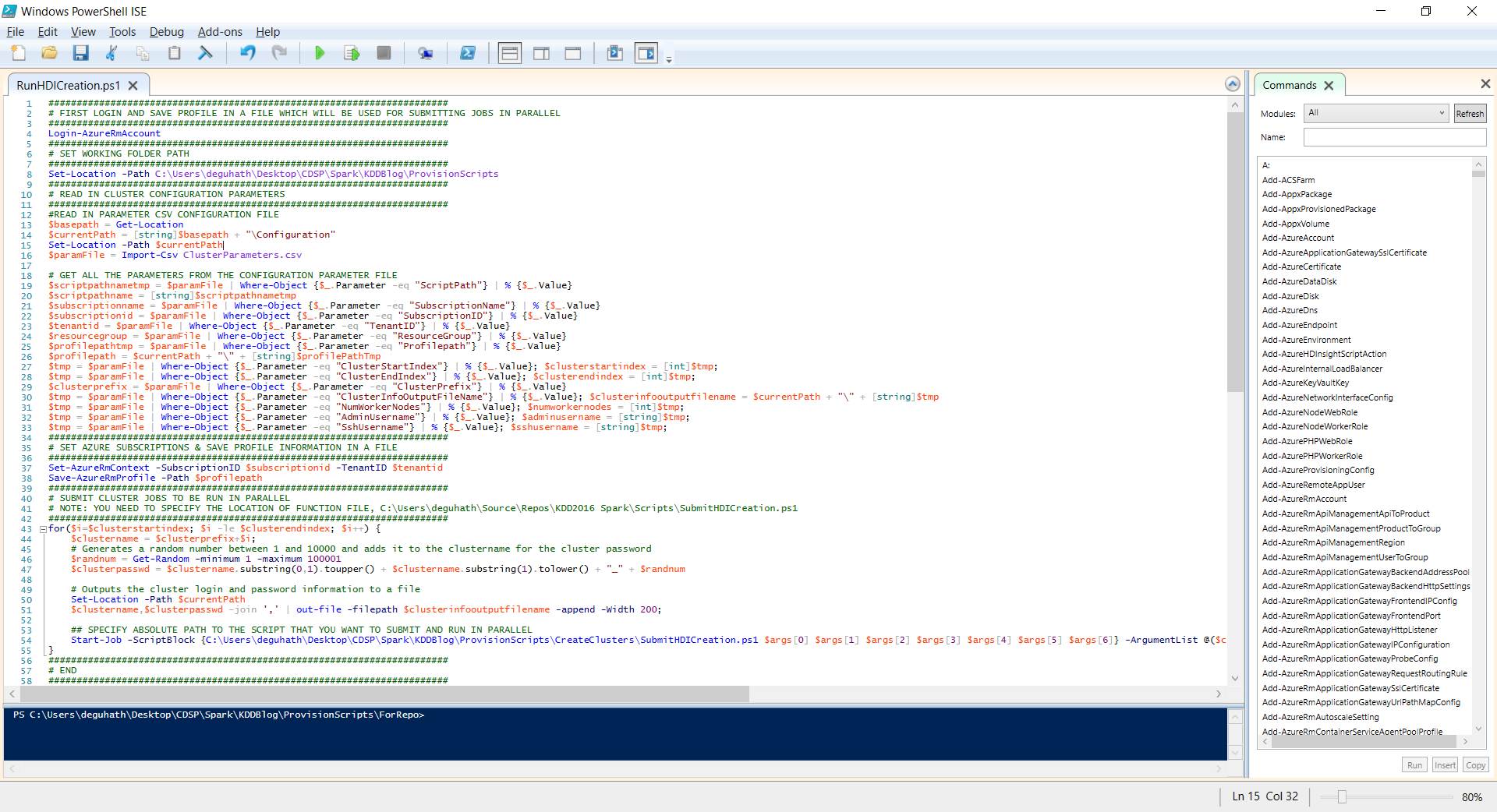
## Creating or Provisioning HDInsight Spark clusters with MRS

An overview of Microsoft R Server on HDInsight (preview) is provided [here](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-r-server-overview/). For provisioning Spark HDInsight clusters using Windows PowerShell, you will need the following configuration files and scripts:

1. A configuration parameter file in csv format to specify clusternames, logins, cluster-size (how many nodes are needed), node-size (size of the VMs to be used as head, edge and worker nodes), Azure subscription which is to be used for creating the clusters, etc. Windows PowerShell will use this as input for getting all the relevant parameters about the clusters. File name: [ClusterParameters.csv](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/ClusterConfigurationParameters.csv)
2. An Azure resource deployment schema in json format for creating the clusters, also called an Azure resource management (ARM) template. The parameters from the CSV configuration file are passed along to this schema file to be used for cluster creation. File name: [azuredeploy.json](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/azuredeploy.json)
3. A Windows PowerShell script which submits the job for creating individual clusters. File name: [SubmitHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/SubmitHDICreation.ps1)
4. A driver Windows PowerShell file which runs the jobs in parallel using the submit script (#4). File name: [RunHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/RunHDICreation.ps1)

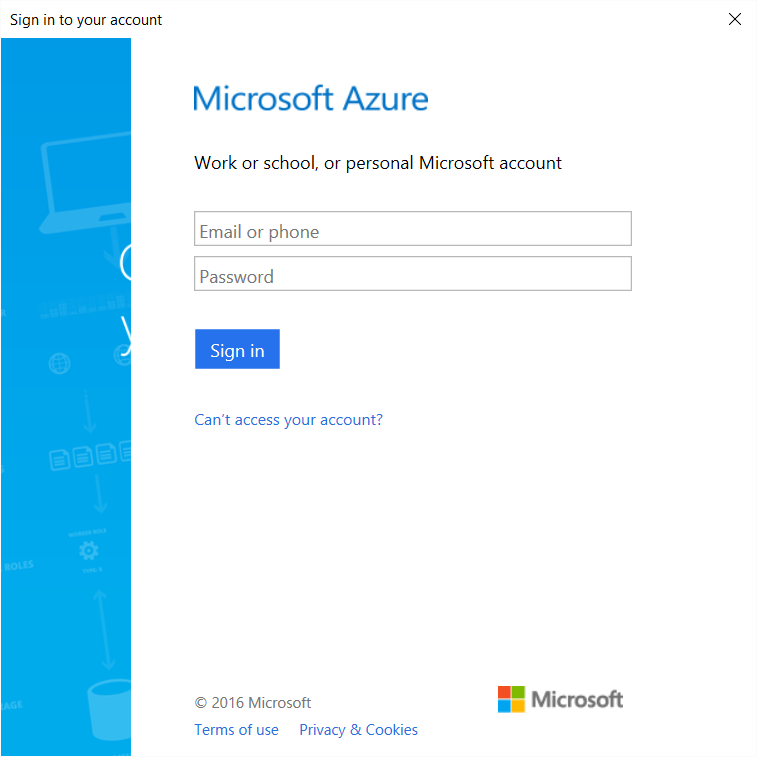
In the public GitHub repository, first two files are in “Scripts/Configuration”, and 3rd and 4th files are in “Scripts/CreateClusters”.

You can copy and paste the code in file: [RunHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/RunHDICreation.ps1). To create clusters, open RunHDICreation.ps1. Set the base path as well as path to [SubmitHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/SubmitHDICreation.ps1). See below in Figure 1.



**Figure 1:** Script “[RunHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/RunHDICreation.ps1)”, used to submit cluster creation jobs. Note that when running the script, you will have to set the paths to the configuration file and script [SubmitHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/SubmitHDICreation.ps1) (shown in red boxes).

The first step is to login to your Azure subscription using the command “Login-AzureRmAccount” in the above file. When you do that, you will be prompted for login and password. You will need to save your profile information in a file, and then pass that profile to [SubmitHDICreation.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/CreateClusters/SubmitHDICreation.ps1) for submitting the batch jobs.



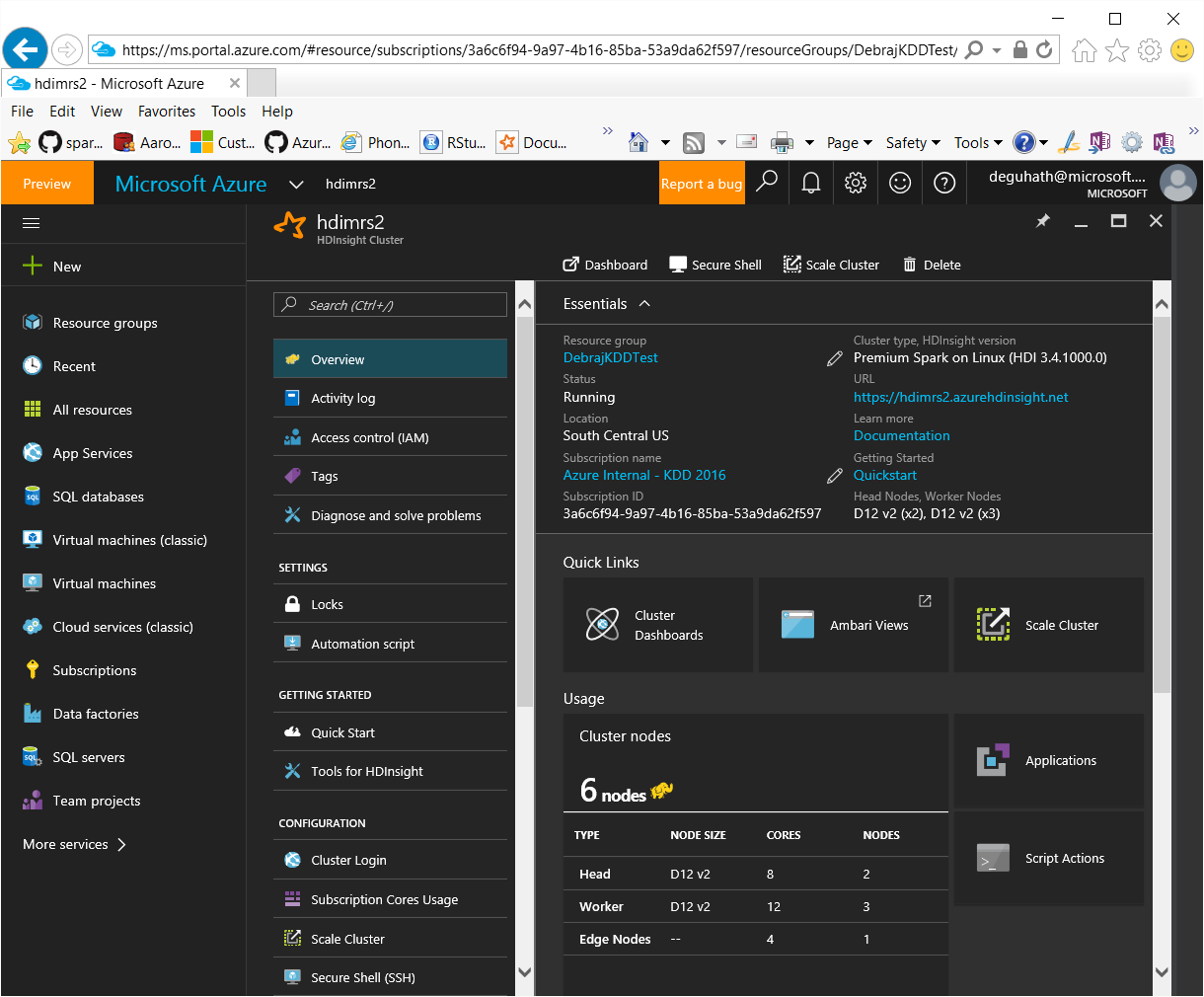
**Figure 2:** Pop-up window for logging into to Azure subscription and getting profile information.

Note that the clusters are currently set to be created using the VMs “Standard\_D12\_v2”, as specified by the following line in the file [azuredeploy.json](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/azuredeploy.json).

"vmSize": "Standard\_D12\_v2"

If you want to change the vmSize – please set the vmSize of your choice in the json file. Details about VM sizes and prices are given [here](https://azure.microsoft.com/en-us/pricing/details/hdinsight/).

Once submitted, the clusters take about 20-25 mins to provision and be active for use. A blob storage account is created for every cluster. The name of the blob storage account will be: name\_of\_cluter + “storage”. For example for a cluster hdimrs1, the corresponding blob storage will be hdimrs1storage. This will serve as the HDFS for the cluster. Once the cluster is ready to be used, you can navigate and examine it through the GUI on your Azure portal.



**Figure 3:** Dashboard for managing HDI Spark cluster on Azure portal. Once the cluster is created, you can access this dashboard for your cluster. You can change the size of the cluster and scale it up or down using the “scale cluster” tile.

## Submitting script actions to install software and R packages

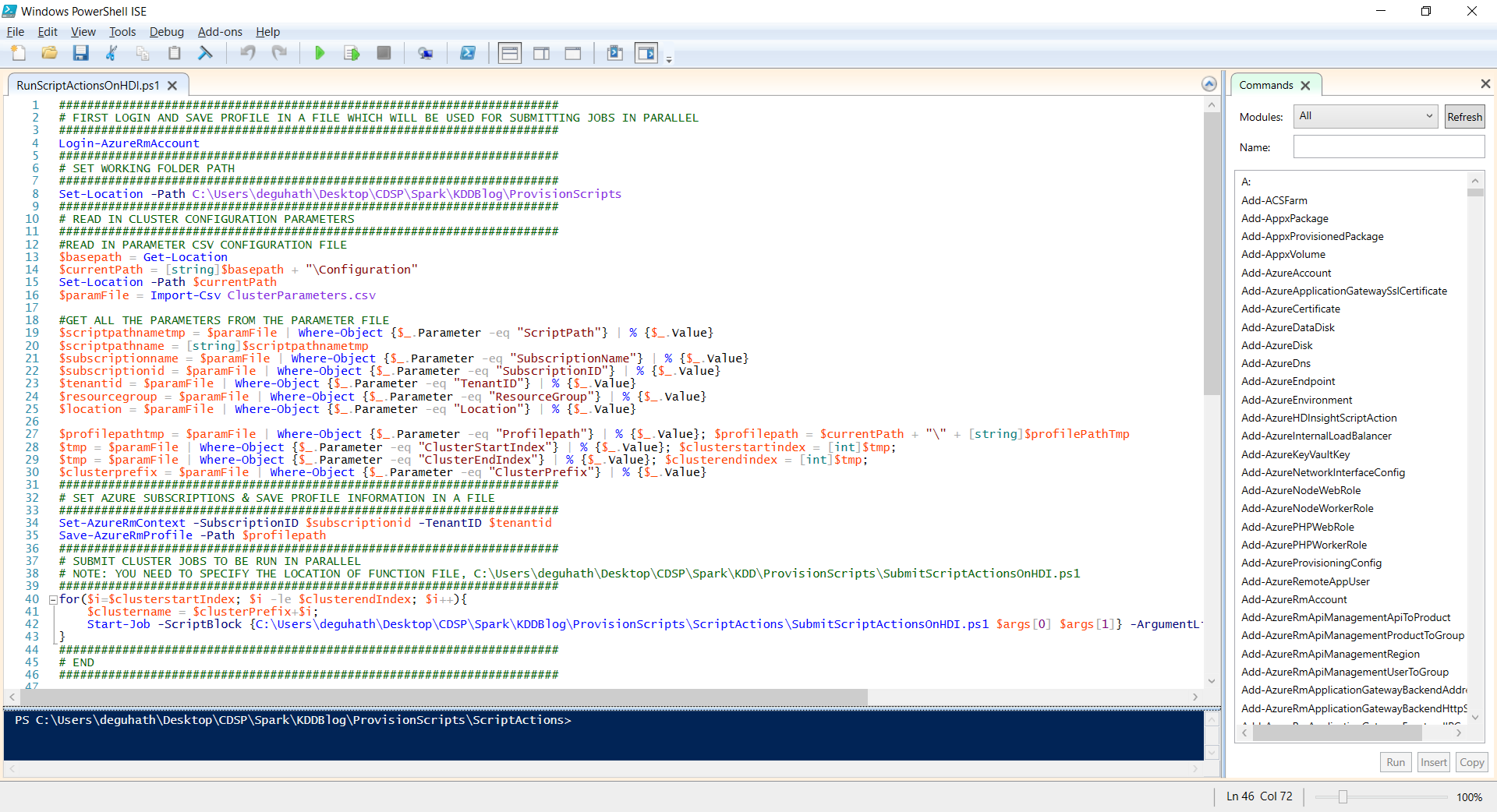
After the cluster is created, you can login to the r-server edge node using a client such as [putty](http://www.putty.org/) or [MobaXterm](http://mobaxterm.mobatek.net/). For example, for a cluster with name hdimrs1, the edge node will have address: r-server.hdimrs2-ssh.azurehdinsight.net. The ssh login and username are needed to login.

Script actions need be run on the edge nodes of the clusters to install necessary software (e.g. RStudio) and R packages, and copy code and script files from the GitHub repository, so that the exercises can run. For more on script actions, read [this](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-customize-cluster-linux/).

For running script actions, you will need the following scripts, which are located in “Scripts/RunningScriptActions”:

1. A parameter configuration file in csv format to specify clusternames, logins, Azure subscription which is to be used for creating the clusters, etc. Windows PowerShell will use this as input for getting all the relevant parameters about the clusters. File name: [ClusterParameters.csv](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/ClusterConfigurationParameters.csv)
2. A Windows PowerShell script which submits the job for creating individual clusters. File name: [SubmitScriptActionsOnHDI.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/SubmitScriptActionsOnHDI.ps1)
3. A driver Windows PowerShell file which runs the jobs in parallel using the submit script (#4). File name: [RunScriptActionsOnHDI.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/RunScriptActionsOnHDI.ps1)

For execution, you can copy and paste scripts from [RunScriptActionsOnHDI.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/RunScriptActionsOnHDI.ps1) on your Windows PowerShell commandline. You will need to specify the location of your configuration file [ClusterParameters.csv] as well as [SubmitScriptActionsOnHDI.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/SubmitScriptActionsOnHDI.ps1). If you have saved your Azure profile in a file earlier, you don’t need to login and save the profile again, you can use that profile.



**Figure 4:** Script [RunScriptActionsOnHDI.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/RunScriptActionsOnHDI.ps1) in Windows PowerShell. Locations where file paths have to be set are indicated in red boxes.

The scripts will download and use these two files from GitHub:

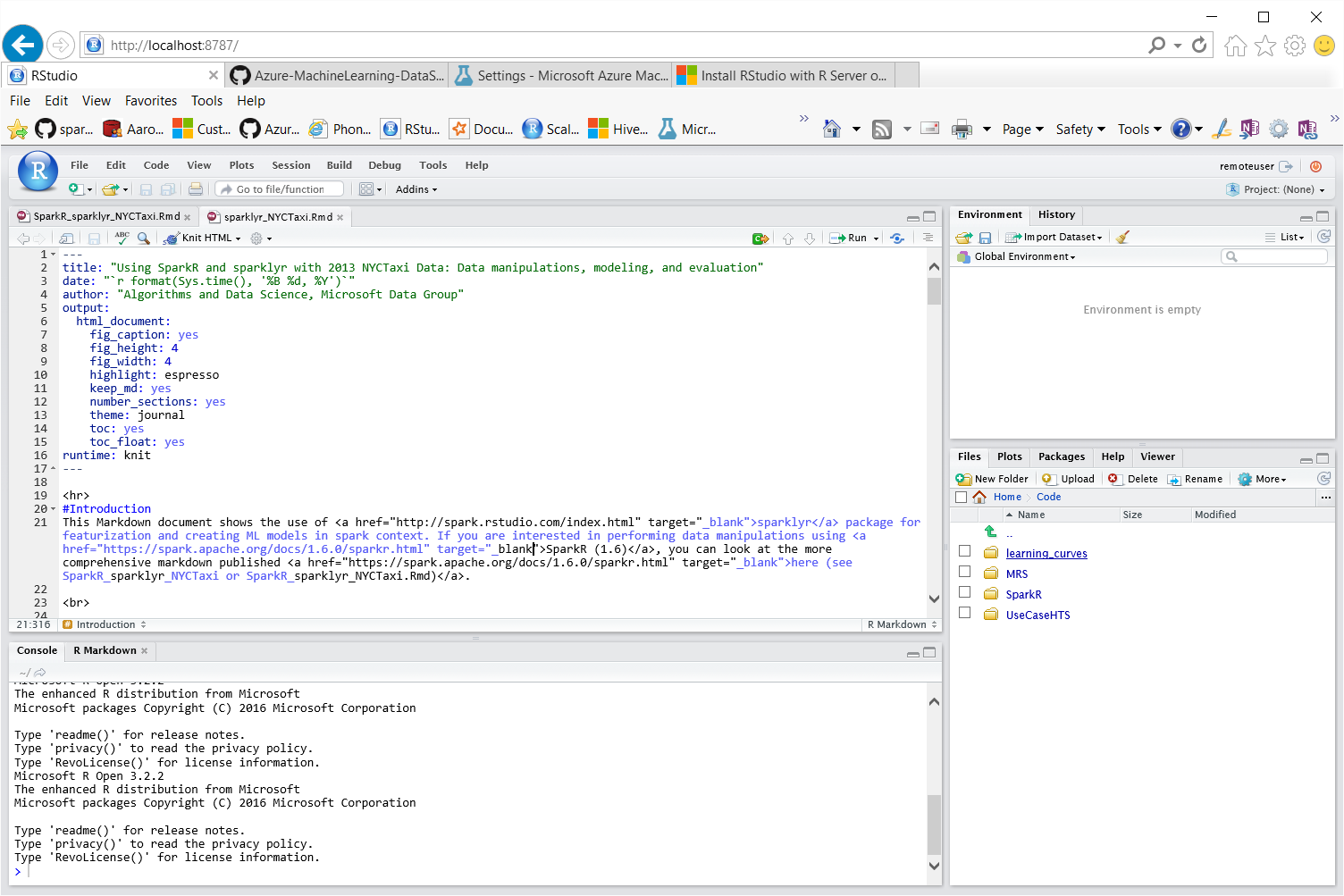
1. An R file for necessary package installations from CRAN and GitHub. File name: [github\_installs.R](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/github_installs.R)
2. A shell script file which calls the R-script file for installing R packages, downloading necessary files and installing RStudio. File name: [downloadRun.sh](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/RunningScriptActions/downloadRun.sh)

## Verifying HDInsight cluster HDFS for code and data files

## Logging into the RStudio server and executing R code in Spark

Instructions for logging into the RStudio server on the edge-node of the Spark cluster are given [here](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-r-server-install-r-studio/#install-rstudio-on-the-cluster-using-a-custom-script). RStudio is already installed on the edge node using script actions. So, you will have to follow instructions on how to login to the RStudio server and run R code. Code files are in GitHub in the “Code” folder.

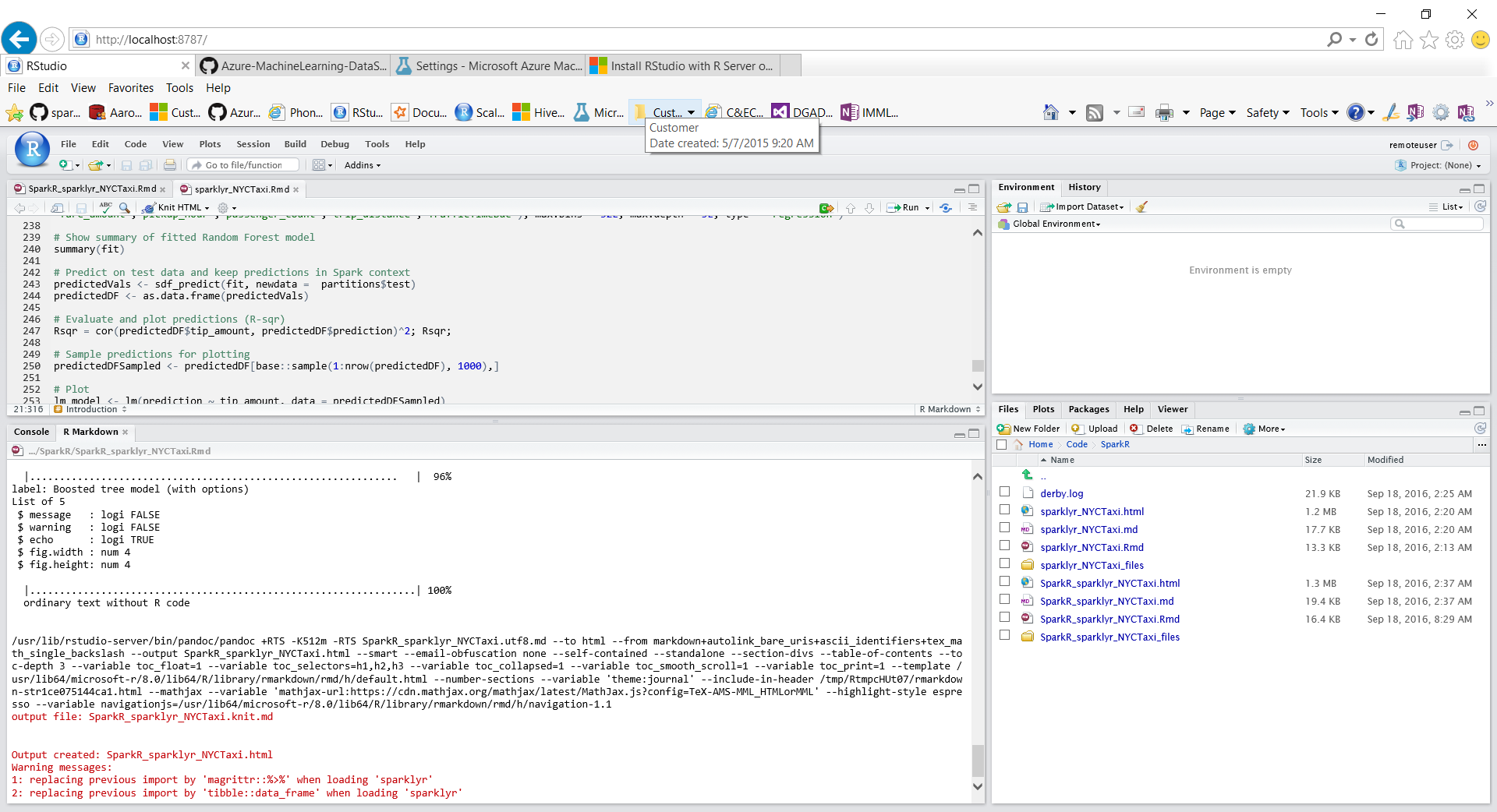
Once you login to the RStudio server, you will see the code files in different directories, as shown below:





**Figure 5:** R markdown file in RStudio. To execute the code, click on “knit HTML”.

For Rmarkdown files (.Rmd), you can click on ‘knit HTML”, and see the code run. After completion of the execution, and HTML output file will be produced. For example, see below.



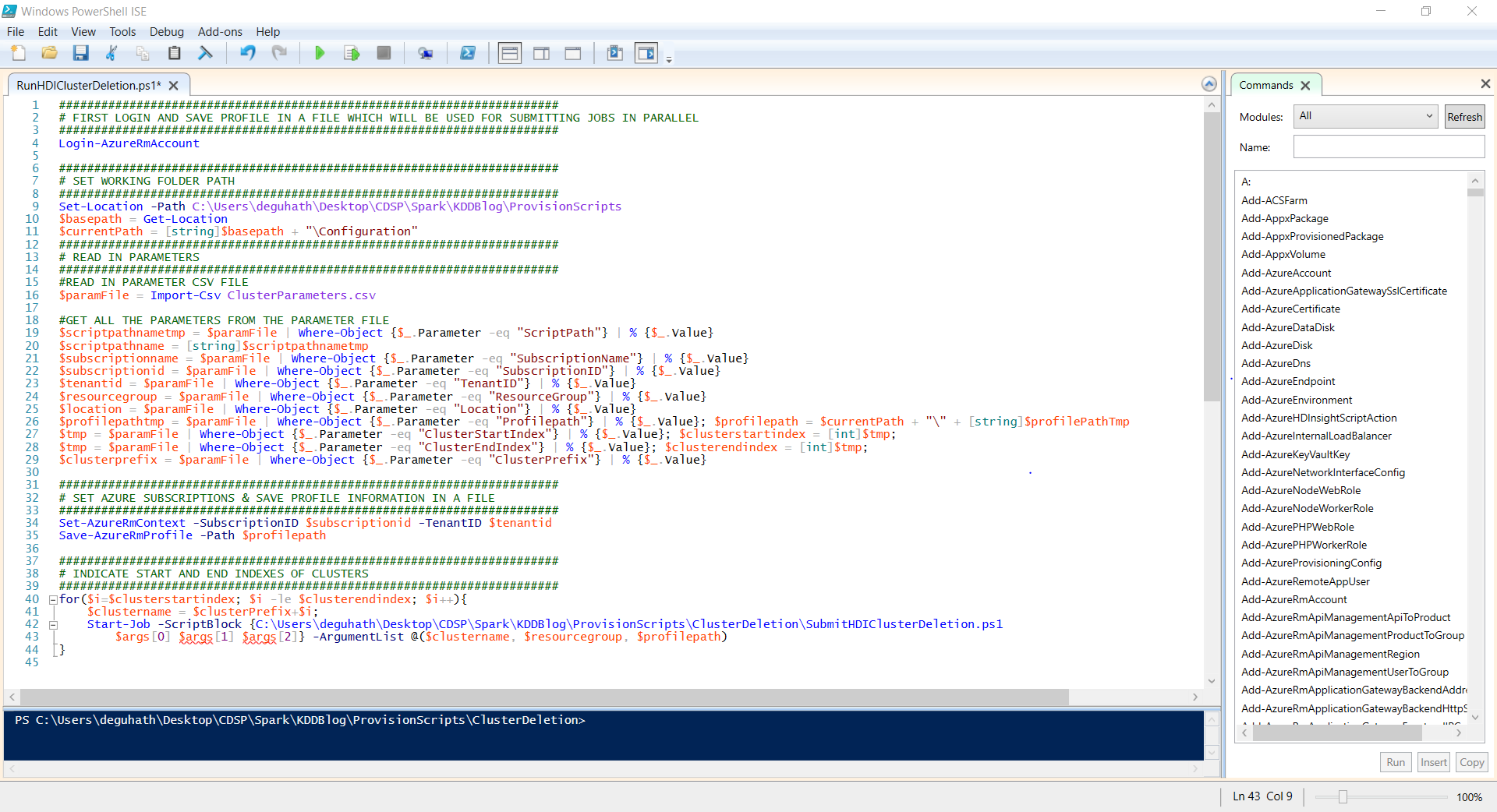
**Figure 6:** RStudio server snapshot after a R markdown file is successfully executed. Note in the bottom right panel the HTML files which appeared after the code in the R markdown file fully executed.

## Deleting HDI clusters

After you are done using the clusters, you can easily delete them using Windows PowerShell scripts and Azure commandlets. For running script actions, you will need the following scripts, which are located in “Scripts/DeleteClusters”.

For cluster deletion, you will need the following files:

1. A parameter configuration file in csv format to specify clusternames, logins, Azure subscription which is to be used for creating the clusters, etc. Windows PowerShell will use this as input for getting all the relevant parameters about the clusters that are to be deleted. File name: [ClusterParameters.csv](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/ClusterConfigurationParameters.csv)
2. A Windows PowerShell script which submits the job for deleting individual clusters. File name: [SubmitHDIClusterDeletion.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/DeleteClusters/SubmitHDIClusterDeletion.ps1)
3. A driver Windows PowerShell script, which reads in the cluster parameters from [ClusterParameters.csv](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/Configuration/ClusterConfigurationParameters.csv), runs the deletion jobs in parallel using the submit script (#1). File name: [RunHDIClusterDeletion.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/DeleteClusters/RunHDIClusterDeletion.ps1)



**Figure 7:** Script [RunHDIClusterDeletion.ps1](https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Misc/KDDCup2016/Scripts/DeleteClusters/RunHDIClusterDeletion.ps1) in Windows PowerShell. Locations where file paths have to be set are indicated in red boxes.