# Azure Data Lake – Lab Manual

# Module 1

## Lab 1 – Account Activation, Resource Groups, and Data Science Virtual Machine Setup

1. Use your existing Azure subscription or get a free trial at <https://azure.microsoft.com/en-us/free/>
2. Learn about Resource Groups - read this page: <https://docs.microsoft.com/en-us/azure/azure-resource-manager/resource-group-portal>
3. Create a Data Science Virtual Machine by following the instructions at <https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/provision-vm>
   1. Locate the “Create your Microsoft Data Science Virtual Machine” section and follow the instructions
   2. Select a size with minimum 2 processors and 6 GB RAM
   3. Use HDD, not SSD
   4. Create a new resource group for the VM
4. Note your admin name and password
5. Log in to the Azure Portal (<http://portal.azure.com>) and start the DSVM
6. Connect to the DSVM and begin updating the Power BI, Visual Studio, and Windows environments
7. Need help? Check here: <https://docs.microsoft.com/en-us/azure/virtual-machines/windows/connect-logon>
8. Don’t forget to turn off the VM after the class

## Lab 2 - Copy and view data on your storage account using AZCOPY

1. Open the Azure Portal, create a Blob Storage Account in the region closest to your DVSM
   1. Note the name and access keys for the Blob Storage Account and copy into Notepad
2. From your DSVM open a command prompt.
3. Locate the Resources.zip file, copy it to the DSVM, and unzip it within the VM
4. Navigate to this page: <https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy>
5. Locate the section marked “Upload blobs to Blob storage” and follow the instructions to load the web log files from the resources.zip in step 3 (all files in iislogs/\*.txt) to your storage account, using your Storage Account and storage keys.
6. Next, locate the section on the web page with instructions marked “Download files from File storage”. Follow the instructions there to copy the files you uploaded in step 5 to any folder on your DSVM.

## Lab 3 - Install and Configure Azure Data Lake Tools

1. Connect to your Data Science Virtual Machine
2. Open Visual Studio
3. Click “Tools” and then “Extensions and Updates”. Ensure “Azure Data Lake Tools” is installed. If not, install it here: <http://aka.ms/adltoolsvs>

# Module 2

## Lab 4 - Set up Azure Data Lake

1. Read all the following instructions before you begin. You can perform these steps from your local workstation or the Data Science Virtual Machine.
2. Navigate to this page, follow all steps you see there: <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-get-started-portal>
3. In step 3, use the Resource Group you created in the previous module.
4. Create a New Data Lake Store with the wizard, leave all defaults.
5. Once you start the deployment, it will take some time to complete.

## Lab 5 - Optional: Secure Data in Azure Data Lake

1. Optional lab: Open this link and follow the instructions you see there: <https://docs.microsoft.com/en-us/azure/active-directory/active-directory-groups-create-azure-portal>
2. NOTE: If you are working with a corporate Azure Account, permissions may prevent you from completing this lab

## Lab 6 - Performance Check using Visual Studio

1. Connect to your Data Science Virtual Machine
2. Open Visual Studio
3. Open this reference and complete all steps: <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-data-lake-tools-get-started>

# Module 3

## Lab 7 - Create and Manage a Data Lake Store

1. We’ll now create a new Data Lake Store, separate from the one we created earlier.
2. Open this page, and follow the instructions under the “Create an Azure Data Lake Store account“ section to create a new data lake account: <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-get-started-portal>
3. Then follow the instructions under the “Create folders in Azure Data Lake Store account” section to create a folder for the web logs we used in module 1.

## Lab 8: Copy files from Azure Blob Storage to the Azure Data Lake Store using AdlCopy

1. Connect to your Data Science Virtual Machine to perform the following lab.
2. Open this reference, and read all the steps you see there: <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-copy-data-azure-storage-blob>
3. Now navigate in the Azure Portal the files you copied to Blob storage in the Second Module. You will use the following steps to copy those four files to the Azure Data Lake Store you created in the last Module.
4. Locate the path, keys, and names of the files you copied to Blob storage and store that information in Notepad.
5. In the Azure Portal, navigate to the Azure Data Lake Store you created in the last Module.
6. Follow the steps in the section marked “Use AdlCopy (as standalone) to copy data from an Azure Storage blob” at the above website for the files in your Blob storage.

## Lab 9: Create a Simple Azure Data Factory

1. Create an Azure Data Factory pipeline which copies the web log files from Azure Blob Storage to Azure Data Lake Store.
2. Seek inspiration in the example you will see here <https://docs.microsoft.com/en-us/azure/data-factory/data-factory-copy-data-wizard-tutorial>
   1. Hint: Use an Azure Data Lake Store as the sink instead of an Azure SQL Database.
   2. See also: <https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-azure-datalake-connector>

## Lab 10: Export data from Azure Data Lake Store

1. Open PowerShell within the Data Science Virtual Machine
2. Follow the “Example” section entry for displaying one of the weblog files you’ve loaded to your Data Lake Store: <https://docs.microsoft.com/en-us/powershell/module/azurerm.datalakestore/Get-AzureRmDataLakeStoreItemContent>
3. Note – you will have to connect to Azure first. The following code will help you do that –

# Sign in to Azure and set the Azure subscription to work   
# with your subscription id by replacing the xxx in the quotes.  
  
$SubscriptionId = "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxx"  
Add-AzureRmAccount  
Set-AzureRmContext -SubscriptionId $SubscriptionId

## Lab 11: Track Azure Data Lake Store Use

1. Open this location, and follow the instructions you see there: <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-diagnostic-logs>

## Lab 12: Create a User for Azure Data Lake Store

1. Open this location and follow all of the steps you see there - <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data>
2. Note: If you are using your an Azure subscription associated with your employer, you might not be able to create security groups in AAD. Instead, choose a security group that you belong to.

## Lab 13: Work with Azure Key Vault

1. Open this location, and follow the steps you see there - <https://docs.microsoft.com/en-us/azure/key-vault/key-vault-get-started>
2. Stop at the section marked “Register an application with Azure Active Directory”.

# Module 4:

## Lab 14: U-SQL in the Azure Portal

1. Run this on the ADLA account you created in the previous modules
2. Open this page and follow all of the steps shown, starting with “Your first U-SQL script” - <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-get-started-portal>

## Lab 15: U-SQL in Visual Studio

1. Connect to your Data Science Virtual Machine for this lab.
2. Open this page and follow all of the steps shown - <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-data-lake-tools-get-started>

## Lab 16: U-SQL using the CLI

1. Connect to your Data Science Virtual Machine for this lab.
2. Open this page and follow all of the steps shown - <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-get-started-cli2>
3. This will create yet another new Data Lake Store and Analytics account.
4. When the tutorial is complete, open your Azure Portal and delete the Data Lake Store and Analytics accounts you just created.
5. Note: Additional Tutorial Labs are here: <https://saveenr.gitbooks.io/usql-tutorial/content/>

## Lab 17: Working with Local Data

1. Connect to your Data Science Virtual Machine for this lab.
2. Open Visual Studio
3. Open this page and follow all of the steps shown - <https://saveenr.gitbooks.io/usql-tutorial/content/getting-started/your-first-usql-script.html>
4. Note: Additional Tutorial Labs are here: <https://saveenr.gitbooks.io/usql-tutorial/content/>

## Lab 18: Processing Weblog Files

1. Open this page, and follow all steps you see there: <https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-analyze-weblogs>

## Lab 19: Optional: Hive Queries in HDInsight using Data Lake Store

1. Optional: Open this page, follow all steps you find there - <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-visual-studio-tools-get-started>

# Module 5

## Lab 20: Analyzing data in Power BI

1. If you have never used Power BI before, fllow these instructions first and then move on to the next step: <https://powerbi.microsoft.com/en-us/documentation/powerbi-desktop-getting-started/>
2. Upload the following file to the Azure Data Lake Store: <https://github.com/Azure/usql/tree/master/Examples/Samples/Data/AmbulanceData/Drivers.txt>
   1. Note: If you have installed the samples in Azure Data Lake Analytics, the file is already uploaded and can be found under /Samples/Data/
3. Follow this example - <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-power-bi>

## Lab 21: Create and Run an Experiment in Azure ML Studio

1. Open this reference and work through all steps - <https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-create-experiment>

# Even more labs…

## Scalable Data Science with Azure Data Lake: An end-to-end Walkthrough

1. <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/data-lake-walkthrough>

## More hands on labs

1. <https://github.com/Azure/AzureDataLake/tree/master/docs/Hands_on_Labs>