Contents

1 INF99X: Sample Course

- **Download Latest Student Handbook and AllFiles Content**
- **Are you a MCT?** - Have a look at our GitHub User Guide for MCTs
- **Need to manually build the lab instructions?** - Instructions are available in the MicrosoftLearning/Docker-Build repository

1.1 What are we doing?

- To support this course, we will need to make frequent updates to the course content to keep it current with the Azure services used in the course. We are publishing the lab instructions and lab files on GitHub to allow for open contributions between the course authors and MCTs to keep the content current with changes in the Azure platform.

- We hope that this brings a sense of collaboration to the labs like we've never had before - when Azure

changes and you find it first during a live delivery, go ahead and make an enhancement right in the lab source. Help your fellow MCTs.

1.2 How should I use these files relative to the released MOC files?

- The instructor handbook and PowerPoints are still going to be your primary source for teaching the course content.

- These files on GitHub are designed to be used in conjunction with the student handbook, but are in GitHub as a central repository so MCTs and course authors can have a shared source for the latest lab files.

- It will be recommended that for every delivery, trainers check GitHub for any changes that may have been made to support the latest Azure services, and get the latest files for their delivery.

1.3 What about changes to the student handbook?

- We will review the student handbook on a quarterly basis and update through the normal MOC release channels as needed.

1.4 How do I contribute?

- Any MCT can submit a pull request to the code or content in the GitHub repro, Microsoft and the course author will triage and include content and lab code changes as needed.

- You can submit bugs, changes, improvement and ideas. Find a new Azure feature before we have? Submit a new demo!

1.5 Notes

1.5.1 Classroom Materials

1.6 It is strongly recommended that MCTs and Partners access these materials and in turn, provide them separately to students. Pointing students directly to GitHub to access Lab steps as part of an ongoing class will require them to access yet another UI as part of the course, contributing to a confusing experience for the student. An explanation to the student regarding why they are receiving separate Lab instructions can highlight the nature of an always-changing cloud-based interface and platform. Microsoft Learning support for accessing files on GitHub and support for navigation of the GitHub site is limited to MCTs teaching this course only.

1.7 title: Online Hosted Instructions permalink: index.html layout: home

2 Content Directory

The lab exercises are listed below.

2.1 Labs

2.2 {% assign labs = site.pages | where_exp:"page", "page.url contains '/Instructions/Labs'" %} | Module | Lab | | --- | --- | {% for activity in labs %}| {{ activity.lab.module }} | [{{ activity.lab.title }}{% if activity.lab.type %} - {{ activity.lab.type }}{% endif %}](/home/ll/Azure_clone/Azure_new/dp-090-databricks-ml/{{ site.github.url }}{{ activity.url }}) | {% endfor %}

2.3 lab: title: 'Getting Started with Azure Databricks' module: 'Module 1 - Introduction to Azure Databricks'

3 Getting Started with Azure Databricks

Azure Databricks is a fast, easy and collaborative Spark based analytics service. It is used to accelerate big data analytics, artificial intelligence, performant data lakes, interactive data science, machine learning and collaboration. You will discover the Azure Databricks environment and the main topics around it: workspace, cluster, notebook.

To begin, you need to have access to an Azure Databricks workspace with an interactive cluster. If you do not have a workspace and/or the required cluster, follow the instructions below. Otherwise, you can skip to the **Upload data** section below.

3.1 Create Azure Databricks resources

To use Azure Databricks, you first need to deploy an Azure Databricks workspace in your Azure subscription and create a cluster on which you will run notebooks and code. You can then upload the data and notebooks to experiment with in your workspace.

3.1.1 Deploy an Azure Databricks workspace

1. In the Azure portal, create a new **Azure Databricks** resource, specifying the following settings:

   - **Subscription**: *Choose the Azure Subscription in which to deploy the workspace.*
   - **Resource Group**: *Create a new resource group.*
   - **Workspace Name**: *Provide a name for your workspace.*
   - **Region**: *Select a location near you for deployment. For the list of regions supported by Azure Databricks, see Azure services available by region.*
   - **Pricing Tier**: Standard

2. Wait for the workspace to be created. Workspace creation takes a few minutes. During workspace creation, the portal displays the Submitting deployment for Azure Databricks tile on the right side. You may need to scroll right on your dashboard to see the tile. There is also a progress bar displayed near the top of the screen. You can watch either area for progress.

3.1.2 Create a cluster

1. When your Azure Databricks workspace resource has been created, go to it in the portal, and select **Launch Workspace** to open your Databricks workspace in a new tab, signing in if prompted.

2. In the left-hand menu of your Databricks workspace, select **Clusters**, and then select **+ Create Cluster** to add a new cluster with the following configuration:

   - **Name**: *Enter a unique name.*
   - **Cluster Mode**: Single Node
   - **Pool**: None
   - **Databricks Runtime Version**: *Select the **ML** edition of the latest available version of the runtime. Ensure that the version selected:*
     - *Does **not** use a GPU*
     - *Includes Scala > **2.11***
     - *Includes Spark > **3.0***
   - **Terminate after**: 120 minutes of inactivity
   - **Node Type**: Standard_DS3_v2

3. Wait for your cluster to be created, which may take several minutes. The cluster will start automatically, and eventually the spinning *Pending* indicator next to the cluster name will change to a solid green circle to indicate a status of *Running*.

3.1.3 Upload data

1. Download `https://raw.githubusercontent.com/MicrosoftLearning/dp-090-databricks-ml/master/data/nyc-t` to your computer, saving it as **nyc-taxi.csv** in any folder.

2. On the **Data** page in the Databricks Workspace, select the option to **Create Table**.

3. In the **Files** area, select **browse** and then browse to the **nyc-taxi.csv** file you downloaded.

4. After the file is uploaded to the workspace, select **Create Table with UI**. Then select your cluster and select **Preview Table**.

5. Specify the following table attributes, and then select **Create Table**.

   - **Table Name**: nyc_taxi
   - **Create in Database**: default
   - **File Type**: CSV
   - **Column Delimiter**: , *(comma)*
   - **First row is header**: *checked*
   - **Infer schema**: *checked*
   - **Multi-line**: *unchecked*

6. After the table has been created, view it in the workspace.

3.1.4 Import Databricks notebooks

1. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*.

2. In the blade that appears, select the downwards pointing chevron (**v**) next to your name, and select **Import**.

3. On the **Import Notebooks** dialog, import the notebook archive from the following URL, noting that a folder with the archive name is created, containing one or more notebooks:

   - `https://github.com/MicrosoftLearning/dp-090-databricks-ml/raw/master/01%20-%20Introduction%20t`

4. Repeat the previous step to import the following notebook archives, noting that a folder is created for each archive as it is imported.

   - `https://github.com/MicrosoftLearning/dp-090-databricks-ml/raw/master/02%20-%20Training%20and%2`
   - `https://github.com/MicrosoftLearning/dp-090-databricks-ml/raw/master/03%20-%20Managing%20Exper`
   - `https://github.com/MicrosoftLearning/dp-090-databricks-ml/raw/master/04%20-%20Integrating%20Az`

3.2 Explore Azure Databricks

In this exercise, you will discover the Azure Databricks environment.

1. In the **01 - Introduction to Azure Databricks** folder in your workspace, open the **Getting Started with Azure Databricks** notebook.

2. In the top left dropdown menu, choose your cluster to attach your notebook to that cluster. *(Alternatively, you will be prompted to attach a cluster when running the first cell in an unattached notebook).*

3. Read the notes in the notebook, running each code cell in turn.

3.3 Clean-up

3.4 If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

3.5 lab: title: 'Working with Data in Azure Databricks' module: 'Module 1 - Introduction to Azure Databricks'

4 Working with Data in Azure Databricks

You will learn to load data by using DBFS and manipulate it by using Spark Dataframes. Databricks File System (DBFS) is a distributed file system mounted into a Databricks workspace and available on Databricks clusters. DataFrames are the distributed collections of data allowing the processing of huge amounts of data.

4.1 Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

4.2 Work with Data in Azure Databricks

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **01 - Introduction to Azure Databricks**, open the **Working with data in Azure Databricks** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

4.3 Clean-up

4.4 If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

4.5 lab: title: 'Preparing Data for Machine Learning' module: 'Module 2 - Training and Evaluating Machine Learning Models'

5 Preparing Data for Machine Learning

Machine Learning is primarily about training models that you can use to provide predictive services to applications. In this exercise, you'll see how you can use Azure Databricks to prepare training data for machine learning purposes.

5.1 Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

5.2 Prepare Data for Machine Learning

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **02 - Training and Evaluating Machine Learning Models**, open the **1.0 Featurization** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

5.3 Clean-up

5.4 If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

5.5 lab: title: 'Training and Validating a Machine Learning Model' module: 'Module 2 - Training and Evaluating Machine Learning Models'

6 Training and Validating a Machine Learning Model

Machine Learning is primarily about training models that you can use to provide predictive services to applications. In this exercise, you'll see how you can use Azure Databricks to train and validate a machine learning model.

6.1 Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

6.2 Train and Validate a Machine Learning Model

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **02 - Training and Evaluating Machine Learning Models**, open the **2.0 Train and Validate ML Model** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

6.3 Clean-up

6.4 If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

6.5 lab: title: 'Using MLflow to Track Experiments' module: 'Module 3 - Managing Experiments and Models'

7 Using MLflow to Track Experiments

MLflow is a fully-featured model tracking and registry system. In this exercise, you will use MLflow to collect model training artifacts, metrics, and parameters. You will then be able to view those outputs through the Azure Databricks UI or programmatically. To begin, you need to have access to an Azure Databricks workspace with an interactive cluster. If you do not have a workspace and/or the required cluster, follow the instructions below. Otherwise, you can skip to the section Upload the Databricks notebook archive.

7.1 Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

7.2 Use MLflow to Track Experiments

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **03 - Managing Experiments and Models**, open the **01 - Using MLflow to Track Experiments** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

   Tip: In the section entitled **View the Experiment, Runs, and Run Details with the Databricks UI**, there will be instructions on what actions to perform, as these actions will take place outside of the confines of a notebook. It may be easiest to open up a second tab in your browser and perform these actions in that tab while reviewing the instructions in the notebook.

7.3 Clean-up

7.4 If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

7.5 lab: title: 'Managing Models' module: 'Module 3 - Managing Experiments and Models'

8 Managing Models

The Azure Databricks model registry is a powerful tool for model registration, model versioning, and tagging models for deployment. In this exercise, you will learn how to use the model registry, through both the User Interface as well as the MLflow API. To begin, you need to have access to an Azure Databricks workspace with an interactive cluster. If you do not have a workspace and/or the required cluster, follow the instructions below. Otherwise, you can skip to the section Upload the Databricks notebook archive.

8.1 Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

8.2 Train and Validate a Machine Learning ModelManage Models

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **03 - Managing Experiments and Models**, open the **02 - Managing Models** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

   Tip: For the first section, **Managing a Model via the User Interface**, there will be instructions on what actions to perform, as many of these actions will take place outside of the confines of a notebook. It may be easiest to open up a second tab in your browser and perform these actions in that tab while reviewing the instructions in the notebook.

8.3   Clean-up

8.4   If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select Terminate to shut it down. Otherwise, leave it running for the next exercise.

8.5   lab: title: 'Running experiments in Azure Machine Learning' module: 'Module 4 - Integrating Azure Databricks and Azure Machine Learning'

9   Running experiments in Azure Machine Learning

Machine Learning is primarily about training models that you can use to provide predictive services to applications. In this exercise, you will learn to run experiments in Azure Machine Learning from Azure Databricks.

9.1   Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** lab to set up your Azure Databricks environment and import the data and notebooks you require.

9.2   Install libraries on the Azure Databricks Cluster

The notebooks you will run depends on certain Python libraries that will need to be installed in your cluster. The following steps walk you through adding these dependencies.

- From within the Azure Databricks workspace, from the **Clusters** section, select your cluster. Make sure the state of the cluster is **Running**.
- Select the **Libraries** link and then select **Install New**.
- In the Library Source, select **PyPi** and in the **Package** text box type `azureml-sdk[databricks]` and select **Install**.
- Next install `sklearn-pandas==2.1.0`
- Next install `azureml-mlflow`

9.3   Deploy an Azure Machine Learning workspace

1. If you have already created an Azure Machine Learning workspace in your subscription, you can skip to the section Exercise: Running experiments in Azure Machine Learning.

2. In the Azure Portal, create a new resource: **Machine Learning**

3. In the Create Machine Learning Workspace dialog that appears, provide the following values:

    - **Subscription**: Choose your Azure subscription.
    - **Resource group**: Select the resource group in which you deployed your Azure Databricks workspace.
    - **Workspace Name**: aml-ws
    - **Region**: Choose a region closest to you (it is OK if the Azure Databricks Workspace and the Azure Machine Learning Workspace are in different locations).

4. Review and complete the creation of Azure Machine Learning workspace.

9.4   Run an experiment in Azure Machine Learning

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and *your_user_name*. Then in the folder named **04 - Integrating Azure Databricks and Azure Machine Learning**, open the **1.0 Running experiments in Azure Machine Learning** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

9.5   Clean-up

9.6   If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the Clusters page, select your cluster and select   Terminate to shut it down. Otherwise, leave it running for the next exercise.

9.7   lab: title: 'Deploying Models in Azure Machine Learning' module: 'Module 4 - Integrating Azure Databricks and Azure Machine Learning'

10   Deploying Models in Azure Machine Learning

Machine Learning is primarily about training models that you can use to provide predictive services to applications. In this exercise, you will learn to train models in Azure Databricks and then deploy models in Azure Machine Learning.

10.1   Prerequisites

Before starting this lab, complete the **Getting Started with Azure Databricks** and **Running experiments in Azure Machine Learning** lab to set up Azure Databricks and Azure machine Learning environments.

10.2   Deploy a Model in Azure Machine Learning

In this exercise, you will learn how to load and manipulate data inside the Azure Databricks environment.

1. In a web browser, open your Azure Databricks workspace.

2. If your cluster is not running, on the **Clusters** page, select your cluster and use the   **Start** button to start it

3. In the Azure Databricks Workspace, using the command bar on the left, select **Workspace**. Then select **Users**, and   *your_user_name*. Then in the folder named **04 - Integrating Azure Databricks and Azure Machine Learning**, open the **2.0 Deploying Models in Azure Machine Learning** notebook.

4. Attach the notebook to your cluster. Then read the notes in the notebook, running each code cell in turn.

10.3   Clean-up

If you're finished working with Azure Databricks for now, in Azure Databricks workspace, on the **Clusters** page, select your cluster and select   **Terminate** to shut it down.

If you have finished exploring Azure Databricks, you can delete the resource groups in your Azure subscription that contain the Azure Databricks and Azure Machine Learning resources.