

**ELT with Azure Data Bricks**

Hands-on lab step-by-step

Feb 2020

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The names of manufacturers, products, or URLs are provided for informational purposes only and Microsoft makes no representations and warranties, either expressed, implied, or statutory, regarding these manufacturers or the use of the products with any Microsoft technologies. The inclusion of a manufacturer or product does not imply endorsement of Microsoft of the manufacturer or product. Links may be provided to third party sites. Such sites are not under the control of Microsoft and Microsoft is not responsible for the contents of any linked site or any link contained in a linked site, or any changes or updates to such sites. Microsoft is not responsible for webcasting or any other form of transmission received from any linked site. Microsoft is providing these links to you only as a convenience, and the inclusion of any link does not imply endorsement of Microsoft of the site or the products contained therein.

© 2018 Microsoft Corporation. All rights reserved.

Microsoft and the trademarks listed at <https://www.microsoft.com/en-us/legal/intellectualproperty/Trademarks/Usage/General.aspx> are trademarks of the Microsoft group of companies. All other trademarks are property of their respective owners.

Contents

[Azure Databricks Hands-on Lab 1](#_Toc33627407)

[Abstract and learning objectives 1](#_Toc33627408)

[Before the Hands-on Lab: 2](#_Toc33627409)

[Task1: Install Azure Storage Explorer 2](#_Toc33627410)

[Task2: Clone the GitHub repository 4](#_Toc33627411)

[Task3: Deploy Azure Resource Group 5](#_Toc33627412)

[Task4: Deploy an Azure Storage Account as below 6](#_Toc33627413)

[Task5: Deploy Azure Databricks 7](#_Toc33627414)

[Exercise 1: Azure Databricks Fundamentals 9](#_Toc33627415)

[Task1: Open Azure Databricks Workspace 9](#_Toc33627416)

[Task2: Import the Databricks Notebooks 9](#_Toc33627417)

[Exercise2: Follow the notebooks in order 14](#_Toc33627418)

# Azure Databricks Hands-on Lab

## Abstract and learning objectives

In this workshop, you will deploy an Azure Databricks workspace and experiment with access data from Azure Storage in form of CSV or Parquet files and transform this data using PySpark and SparkSQL.

Note: This lab is designed to complement the classroom presentation and the instructor will guide the attendees which exercises should be attempted in which order.

## Before the Hands-on Lab:

**Note on naming conventions:** If you are using a shared environment (subscription) follow this naming convention to make sure your resources are easily identifiable from other participants. ***hol-<your-name>-<resource-name>***

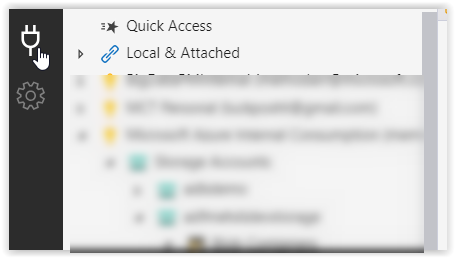
You will be provisioning the below resources:

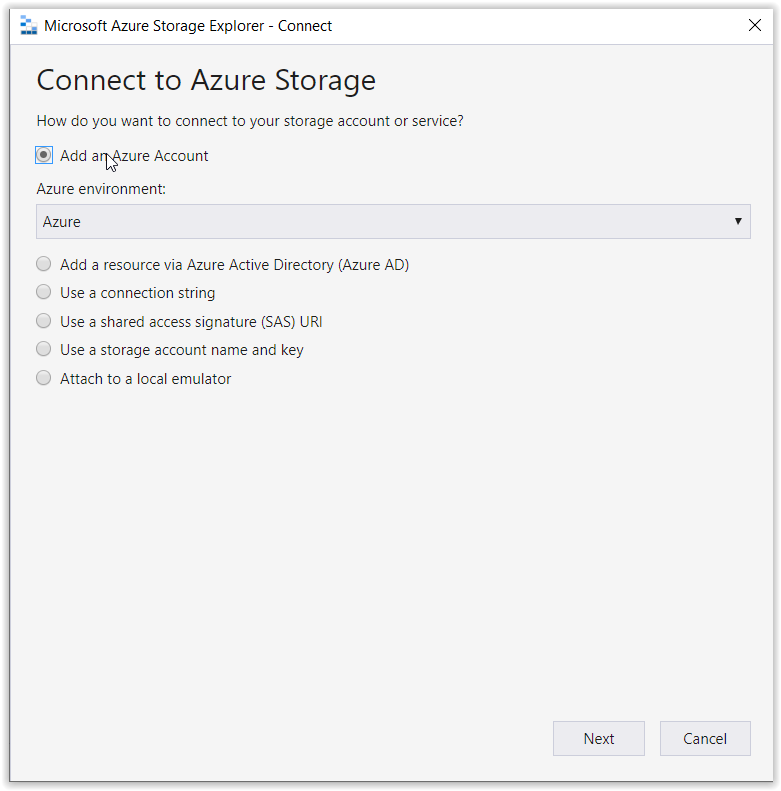
1. Resource Group (If your environment has not already provided this). : *hol-andrew-rg*
2. Azure Storage Account : *holandrewstorage* – For storage name only alphanumeric is allow.
3. Azure Databricks workspace : *hol-andrew-databricks*

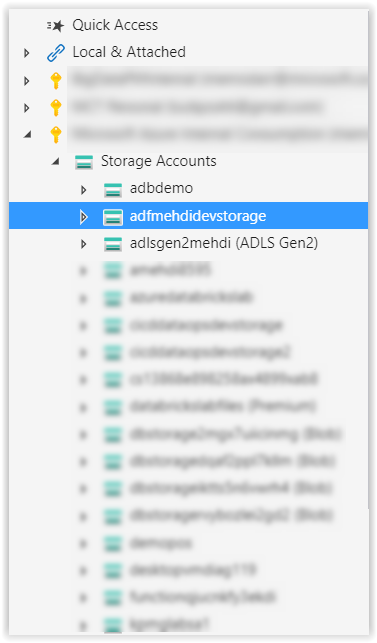
#### Task1: Install Azure Storage Explorer

Microsoft Azure Storage Explorer is a standalone app that makes it easy to work with Azure Storage data on Windows, macOS, and Linux (Azure Docs on Storage Explorer [here](https://docs.microsoft.com/en-us/azure/vs-azure-tools-storage-manage-with-storage-explorer?tabs=windows))

1. Download Azure Storage Explorer from [this page](https://azure.microsoft.com/en-us/features/storage-explorer/)
2. Install Azure Storage Explorer
3. Click the Open Connect dialog
4. Select “Add an Azure account” and continue logging in to your Azure account.
5. Once you log in you can see a list of your storage accounts.

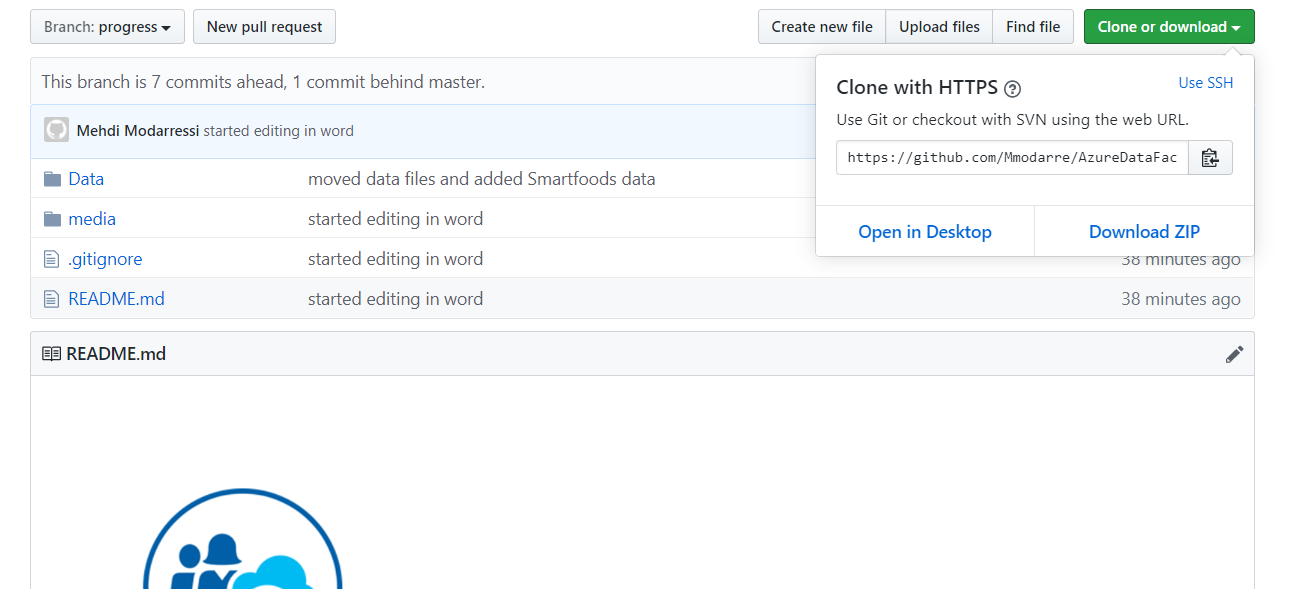






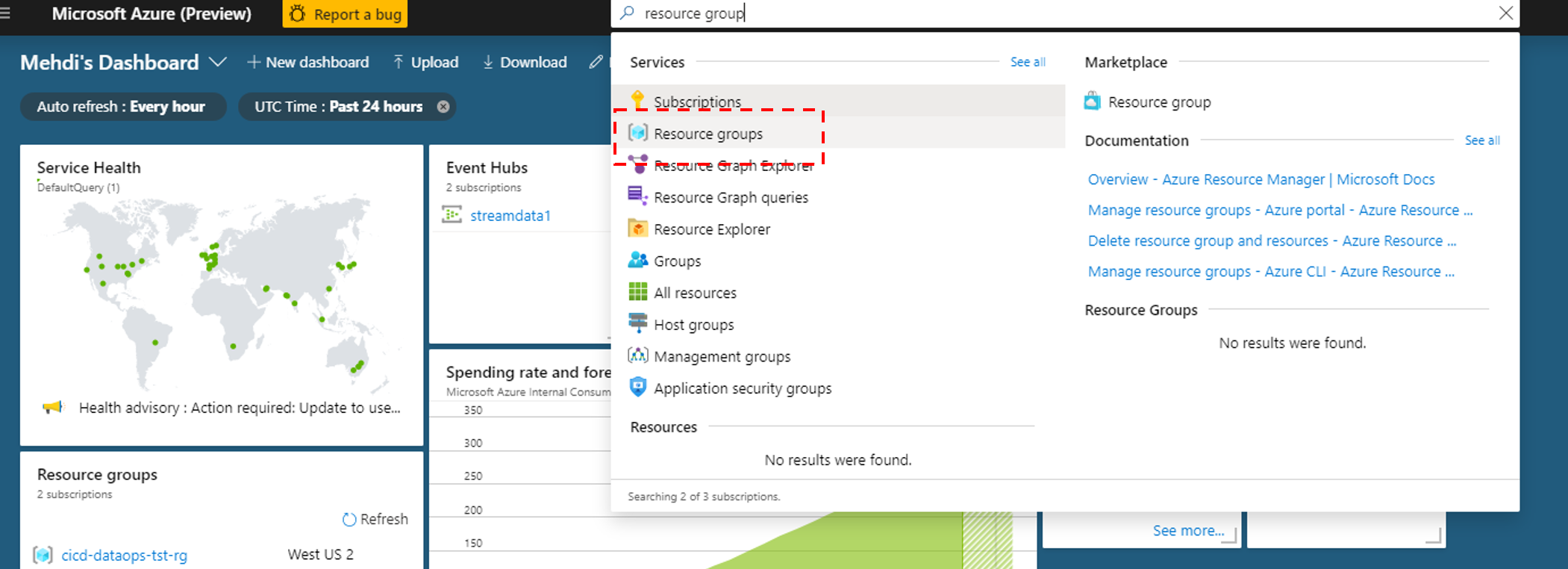
#### Task2: Clone the GitHub repository

Clone this GitHub repository to your machine or download the source Zip file of this repo

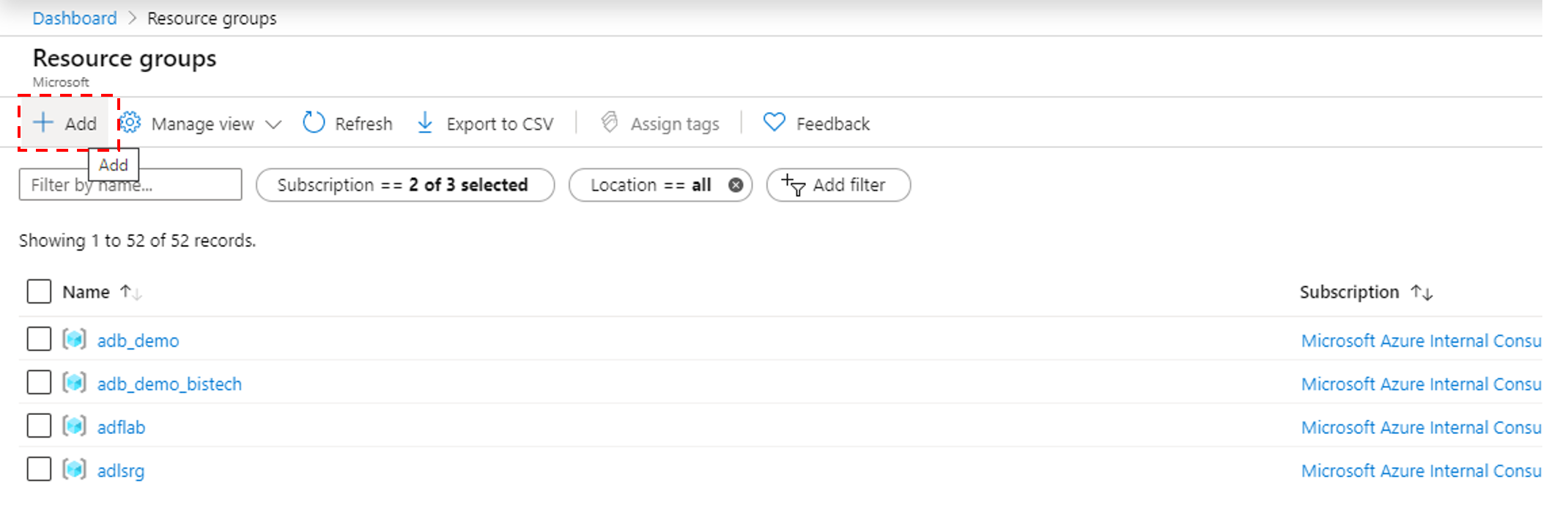


#### Task3: Deploy Azure Resource Group

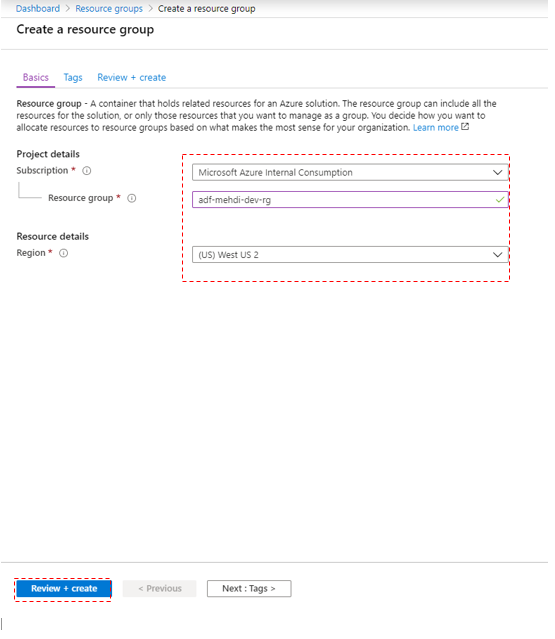
1. Search for Resource groups from the main search bar in Azure portal



1. Click Add

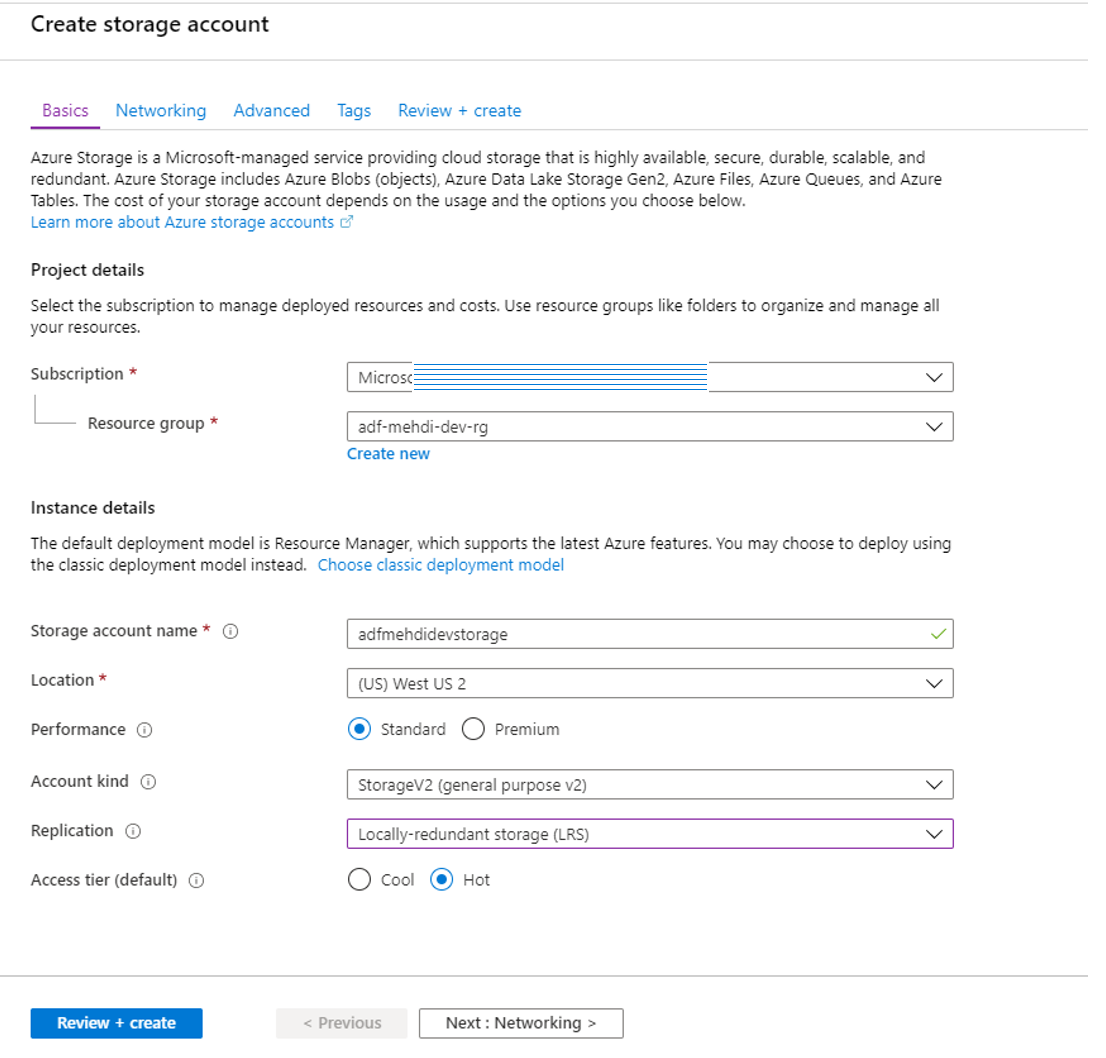


1. **For Resource group name: hol-<your name>-rg for example: hol-john-rg**



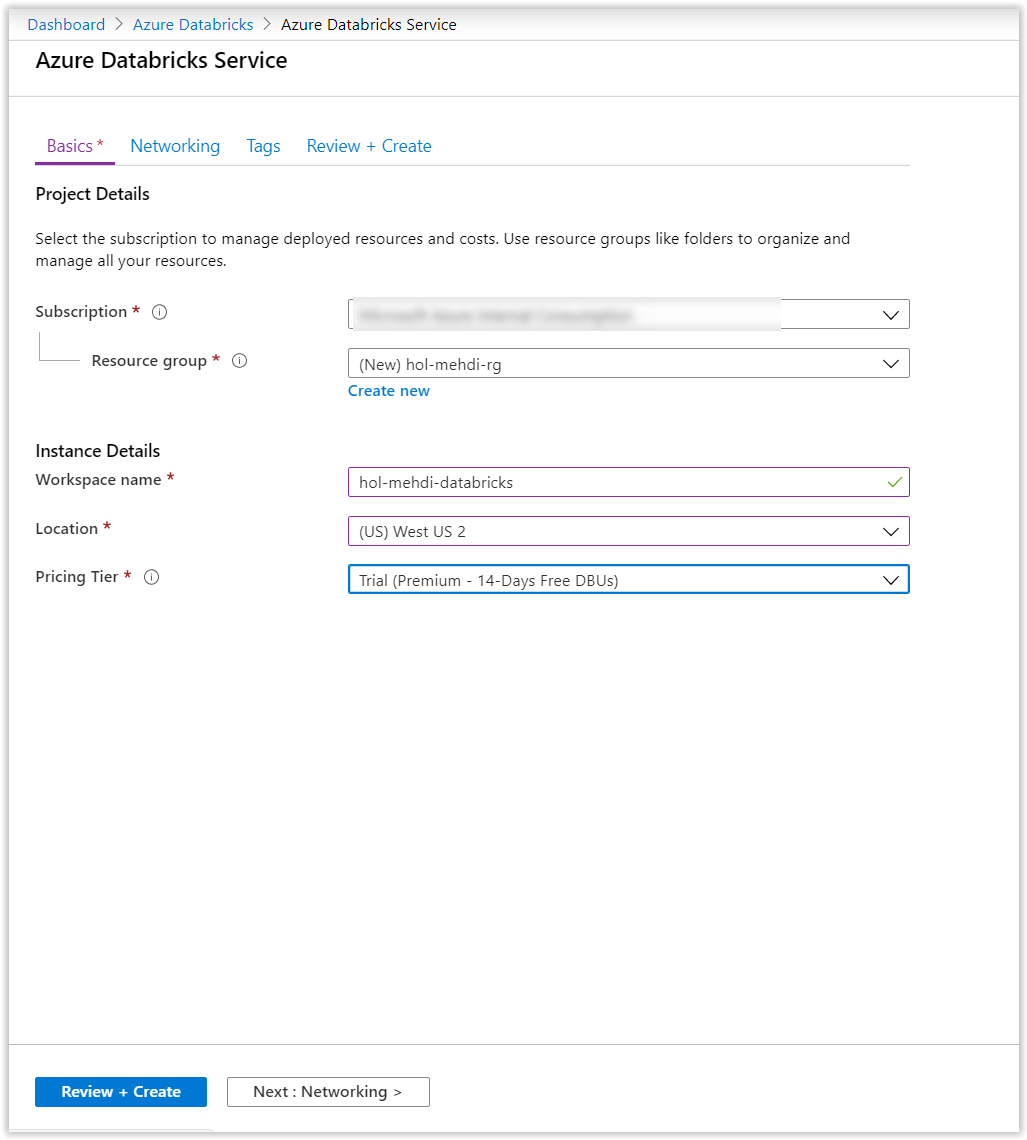
#### Task4: Deploy an Azure Storage Account as below

1. Search for Storage Accounts
2. Click Add
3. Provide subscription, resource group
4. **For Storage account name: hol<your name>storage for example: holjohnstorage**
5. Select Location
6. For Account Kind Storage V2
7. **For Replication change to LRS**
8. Click “Review + Create” and then “Create”



#### Task5: Deploy Azure Databricks

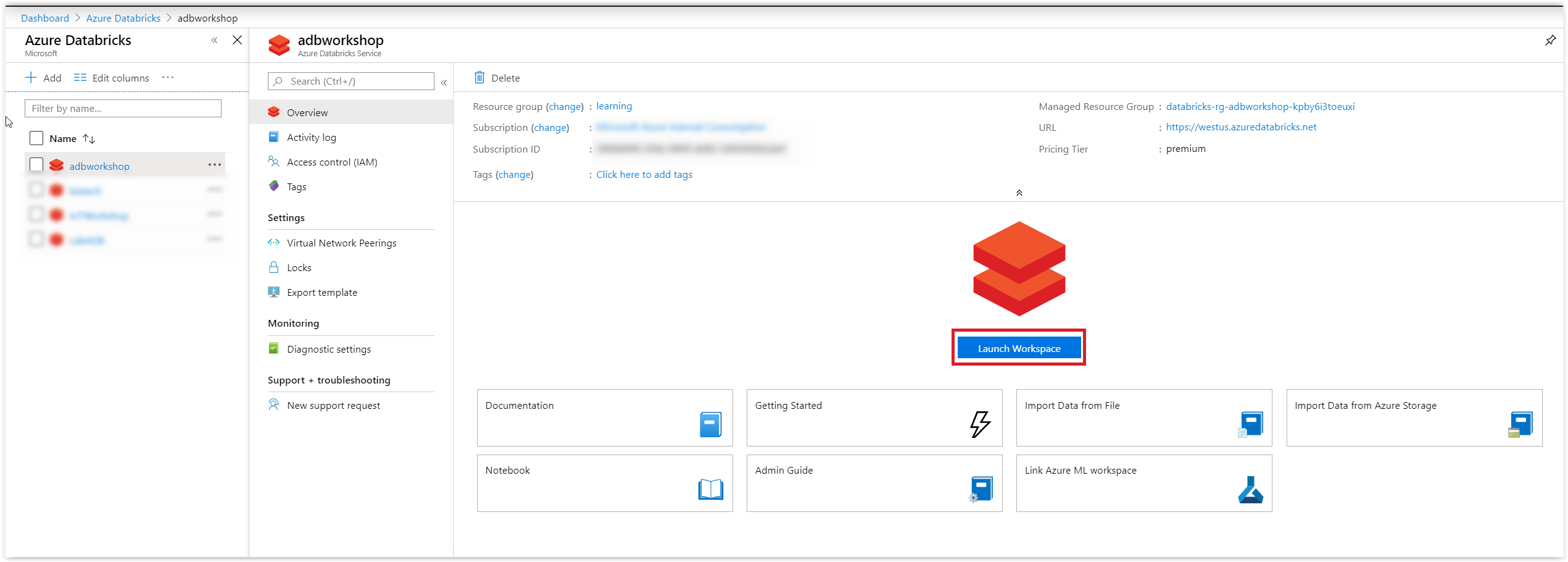
1. Search for “Databricks” in Azure portal
2. Click Add
3. In Creation screen:
   1. Select your subscription
   2. Select your Resource Group
   3. Provide the name hol-<yourname>-databricks
   4. Select Location
   5. Pricing Tier: **Trial**
   6. click Review+Create
   7. Click Create



## Exercise 1: Azure Databricks Fundamentals

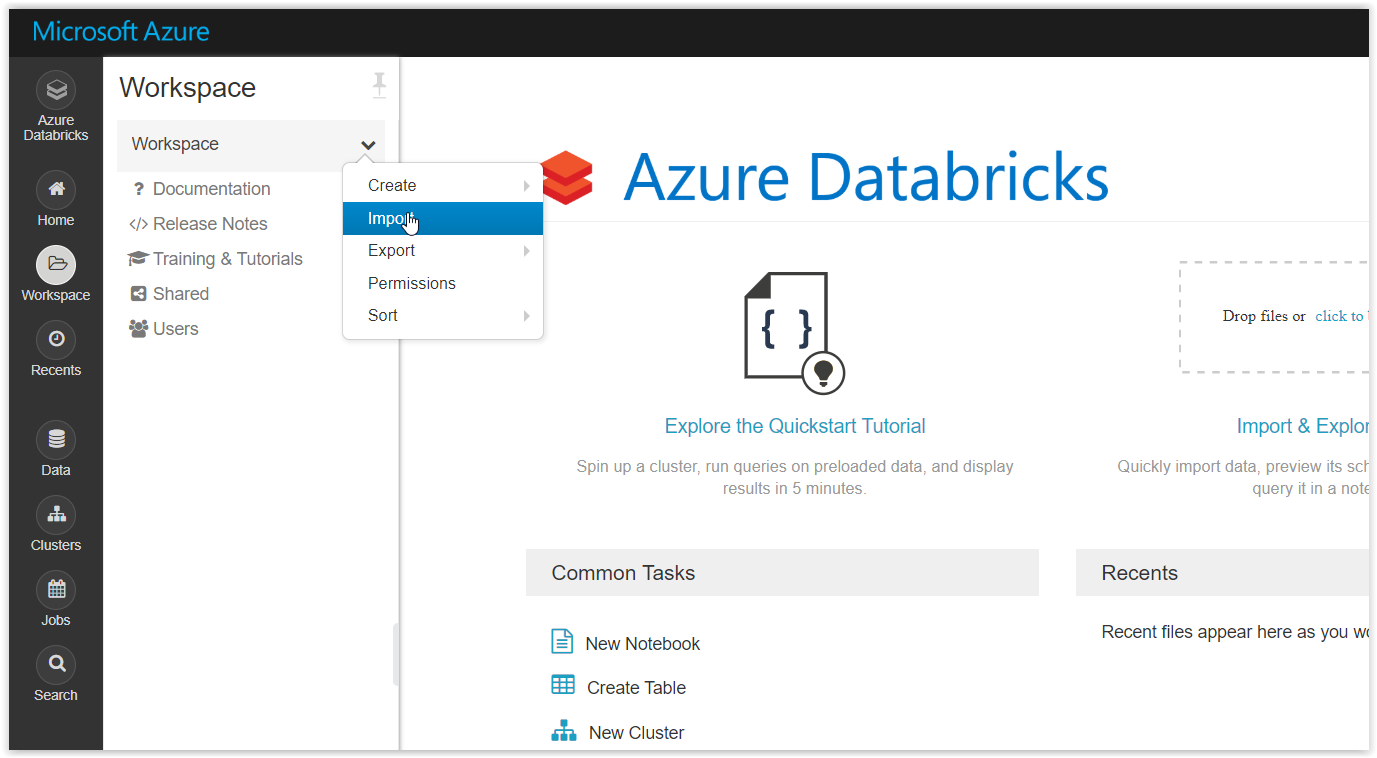
#### Task1: Open Azure Databricks Workspace

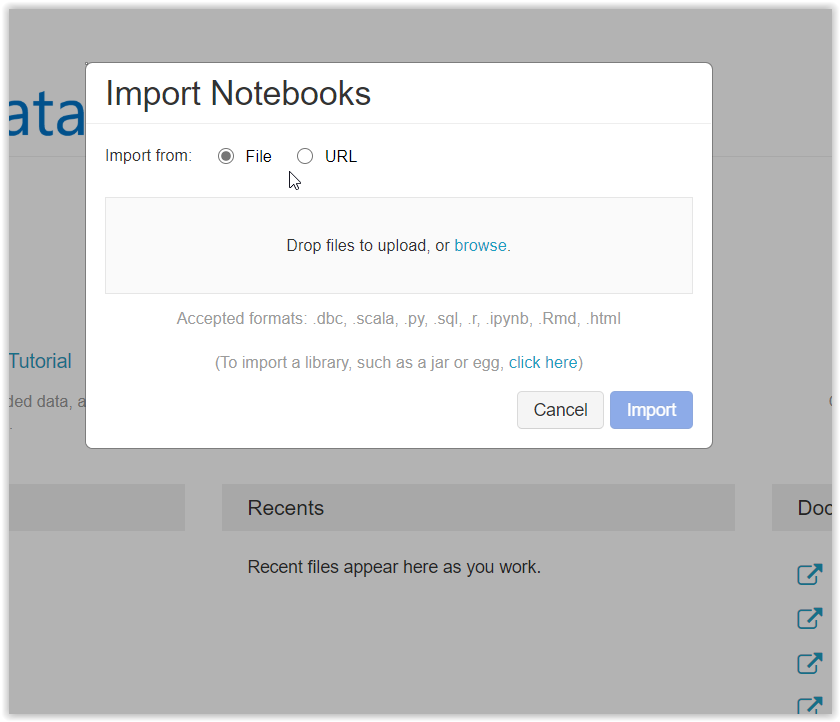
1. Go to the Azure Databricks resource in Azure portal
2. Click “Launch Workspace” to open the Databricks Workspace



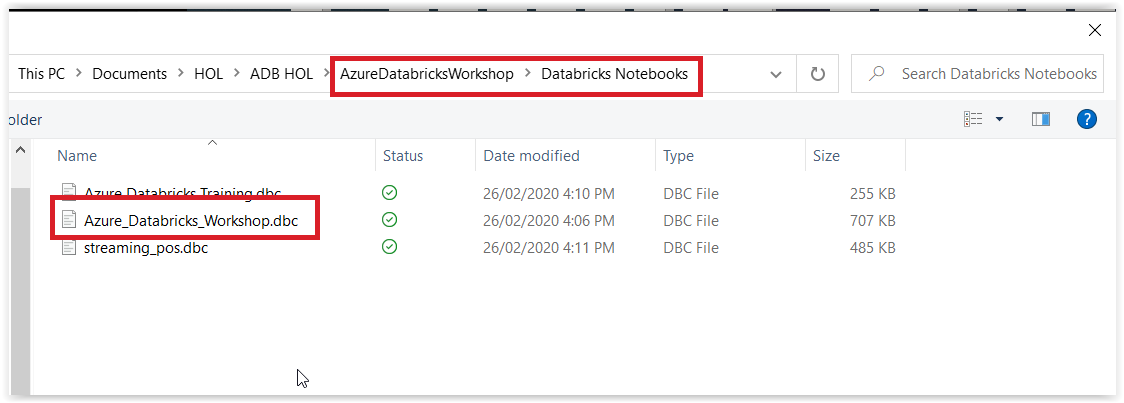
#### Task2: Import the Databricks Notebooks

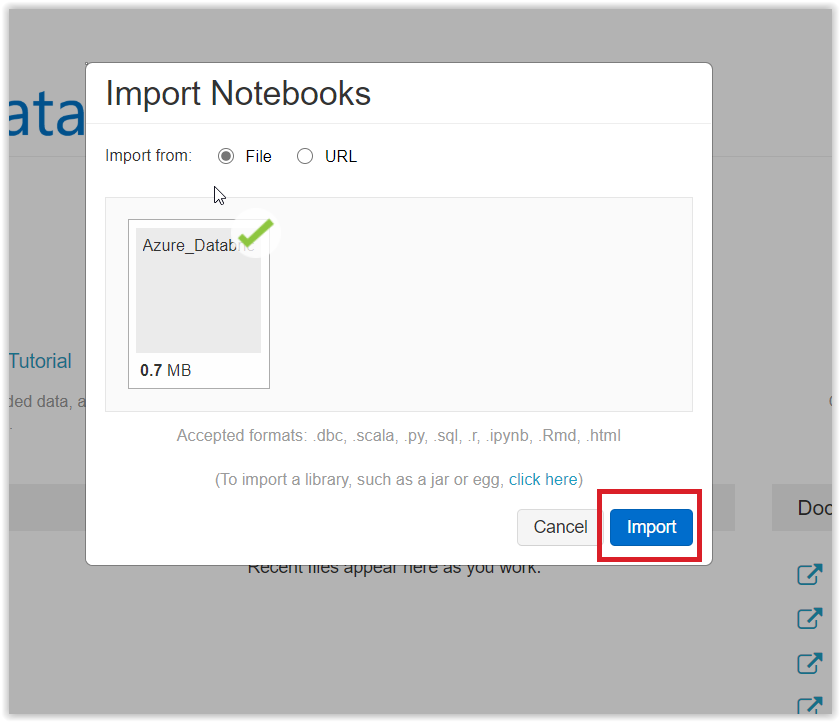
1. From left hand bar select “Workspace”
2. Click the downward arrow next to workspace then and click import



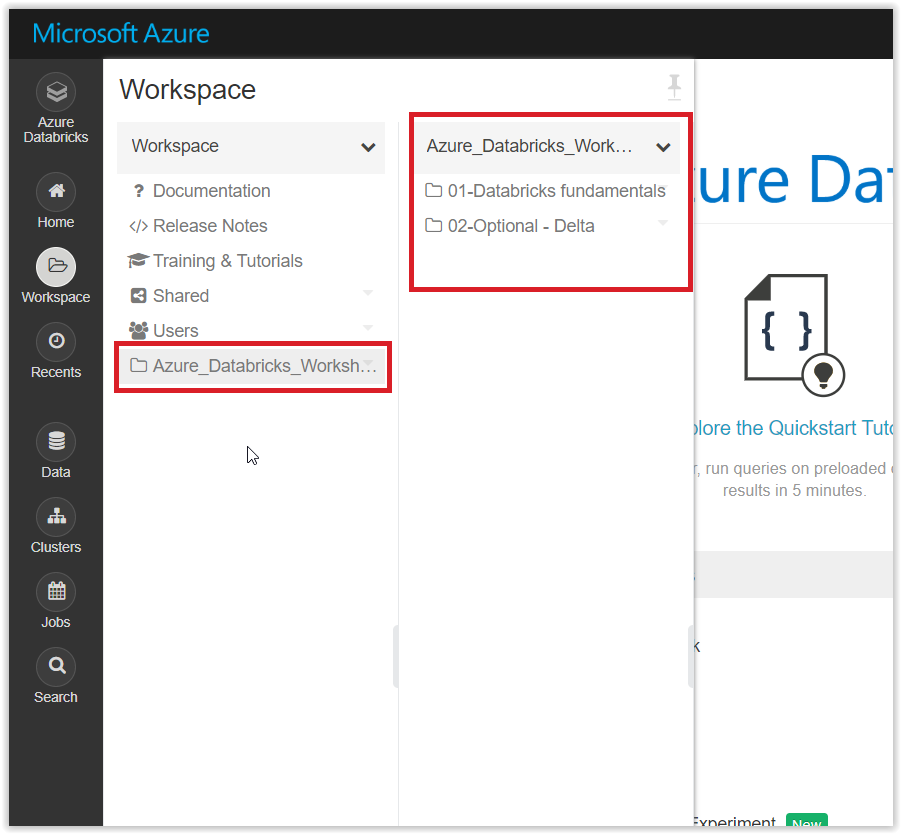


1. Select “File” and Browse to upload the DBC (Databricks Archive file)
2. From the folder that you cloned/downloaded the GitHub repo under “Databricks Notebooks” select “Azure\_Databricks\_Workshop





1. Click “Import”
2. Once Import is done you should be able to see Azure Databricks Workshop directory which has two sub-directories in it.



## Exercise2: Follow the notebooks in order

Start with Notebook 01 under “01-Databricks fundamentals” and go through the notebooks.