Microsoft Cloud Workshops

Data Ingestion and Reporting

Walkthrough

April 2019

© 2017 Microsoft Corporation. All rights reserved. This document is confidential and proprietary to Microsoft. Internal use only. This document is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.

This document is provided "as-is." Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

Some examples are for illustration only and are fictitious. No real association is intended or inferred.

Contents

[Overview 1](#_Toc5120900)

[Objectives 1](#_Toc5120901)

[The Workshop 1](#_Toc5120902)

[Delivery 1](#_Toc5120903)

[Duration 2](#_Toc5120904)

[Requirements 2](#_Toc5120905)

[Workshop Step-by-Step 3](#_Toc5120906)

[Lab 01: Setup 3](#_Toc5120907)

[Exercise 1: Access Azure 3](#_Toc5120908)

[Exercise 2: Access GitHub Repository 4](#_Toc5120909)

[Exercise 3: Create Local File Storage Location 4](#_Toc5120910)

[Lab 02: Create Raw Data Repository In Azure 9](#_Toc5120911)

# Overview

This workshop is designed to guide participants through a common data ingestion and reporting scenario. As we work through this activity, we will explore alternative technologies, architectures and approaches that could be utilized to address different environments and requirements. These include unstructured data, big data, high velocity data and deployment scenarios.

## Objectives

Explore and gain hands on experience with the Azure Data stack of technologies.

Develop a working end to end ingestion, analytics and reporting solution.

Discuss options, architectures and alternative technologies to fit specific requirements.

## The Workshop

The data we will be using is part of the World Development Indicators (WDI) dataset developed by the World Bank. Health data has been extracted from the larger set of datapoints and has been provided in CSV file format in 2 files. One file contains WDI country data and the other file contains yearly indicator records for each country.

We will firstly upload these files into Azure Blob Storage using Azure Storage Explorer. Then we will deploy and secure an Azure SQL Database and create the database artifacts required to store and retrieve the data. Next, we create an Azure Data Factory pipeline that will load our WDI data into the Azure SQL Database artifacts we created earlier. The next step is to connect to this repository using Azure Databricks where we will manipulate a dataset and use it to train a predictive model to provide data insights, we will write these results back to the SQL database. The final step will be to surface these insights and the WDI health data utilizing Power BI in the form of an interactive report.

## Delivery

Our instructors will walk through each exercise and then provide assistance to participants to complete their own implementation.

Please feel free to ask questions of any of us at any time and ask for help when required.

We want to promote an interactive environment and foster 2-way discussions on how and when our technologies and services can be utilized to enable your outcomes, so please feel open to engage and share your ideas.

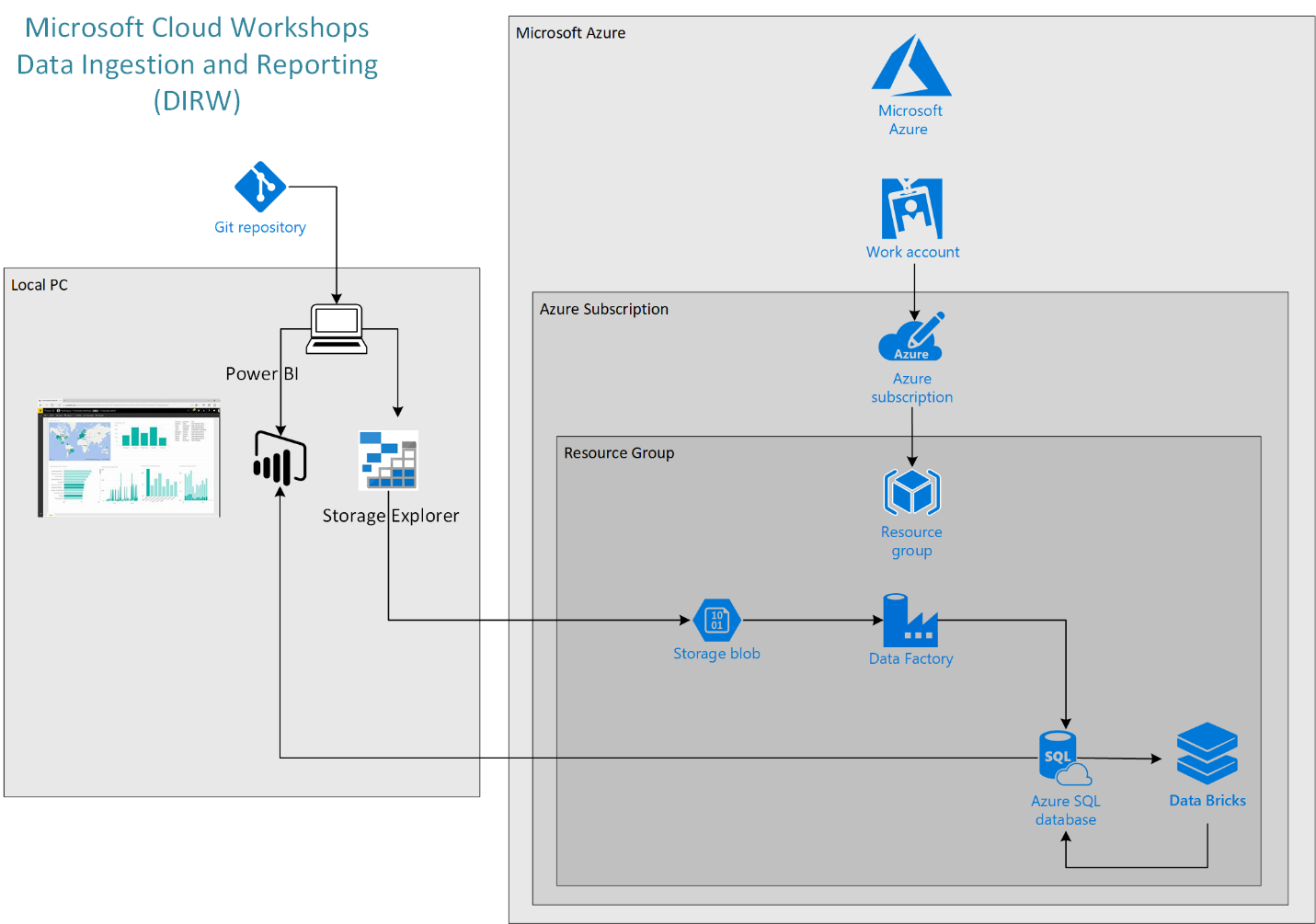
## Duration

4 – 5 Hours.

## Requirements

* Microsoft Azure subscription with permissions to deploy the following services:
  + Azure Blob Storage
  + Azure SQL Server DB
  + Azure Data Factory
  + Azure Databricks
* Laptop with:
  + Azure Storage Explorer installed (latest version: <https://azure.microsoft.com/en-us/features/storage-explorer/>)
  + Power BI Desktop (latest version: <https://powerbi.microsoft.com/en-us/desktop/)>

# Workshop Step-by-Step



## Lab 01: Setup

In this lab we will test access to the required online resources for this workshop and create the local raw data repository.

Prerequisites

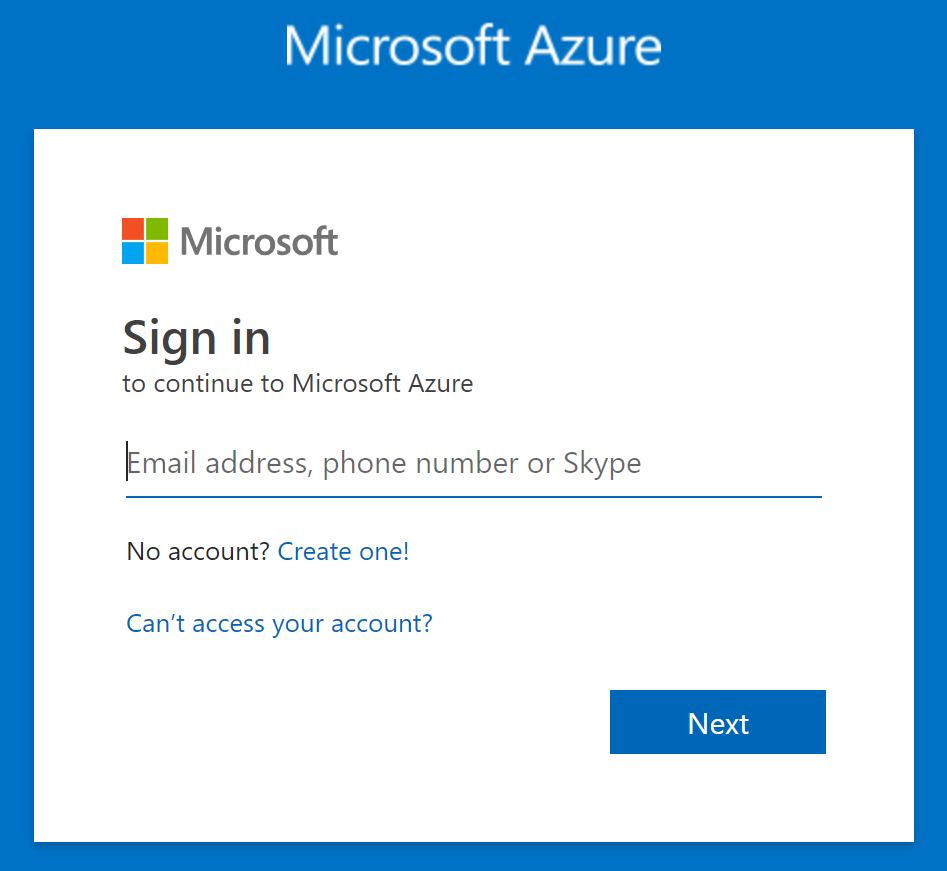
* Microsoft Azure subscription, please have your subscription details available for this workshop. If you do not have your own subscription (recommended), a free trial subscription can be accessed here: <http://azure.microsoft.com/en-us/pricing/free-trial/>

### Exercise 1: Access Azure

This is an Azure based workshop, here we will ensure your subscription is good to go.

*Task 1: Access and bookmark the Azure Portal*

1. Open a web browser
2. Navigate to the [Azure portal](https://portal.azure.com/)
3. Enter your login details:



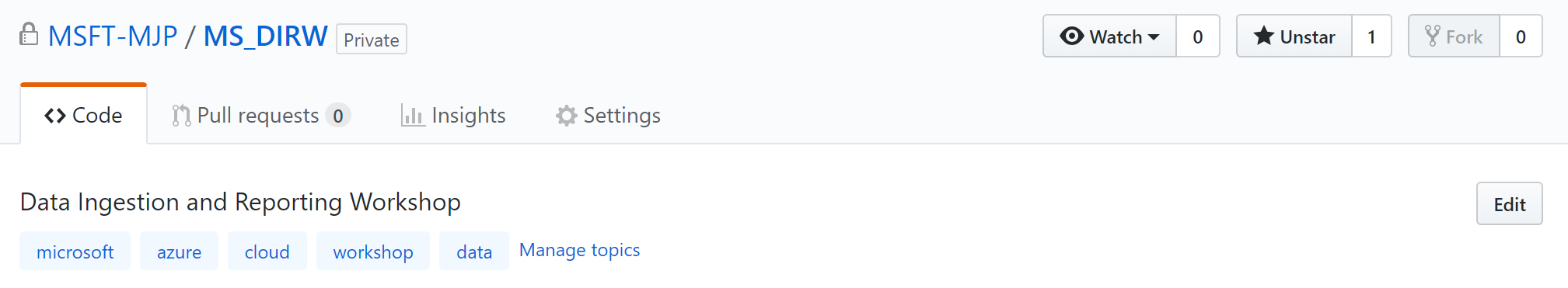
1. You should now see your Azure dashboard, bookmark this page for future reference

### Exercise 2: Access GitHub Repository

All files, documentation and other artifacts used in this workshop have been published to a GitHub repository.

*Task 1: Navigate and bookmark GitHub repository*

1. Open a web browser
2. Navigate to <https://github.com/MSFT-MJP/MS_DIRW>



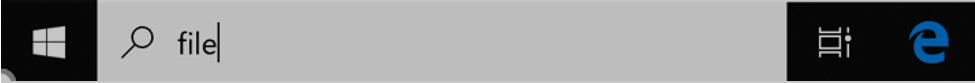
1. Bookmark this repository for future reference

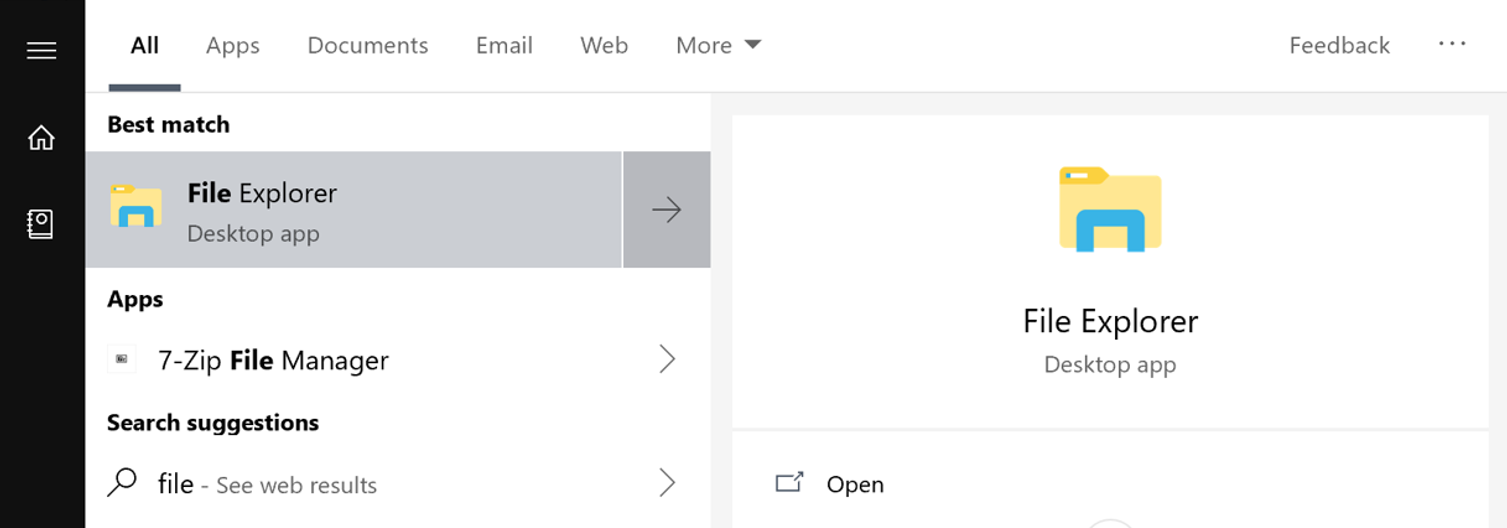
### Exercise 3: Create Local File Storage Location

We will need a local copy of the raw data to be used in this workshop.

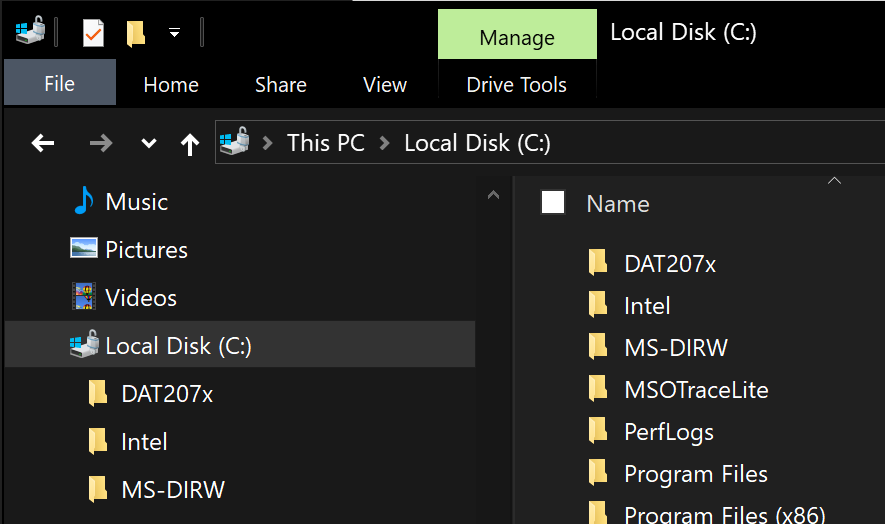
*Task 1: Create local folder*

1. Open File Explorer



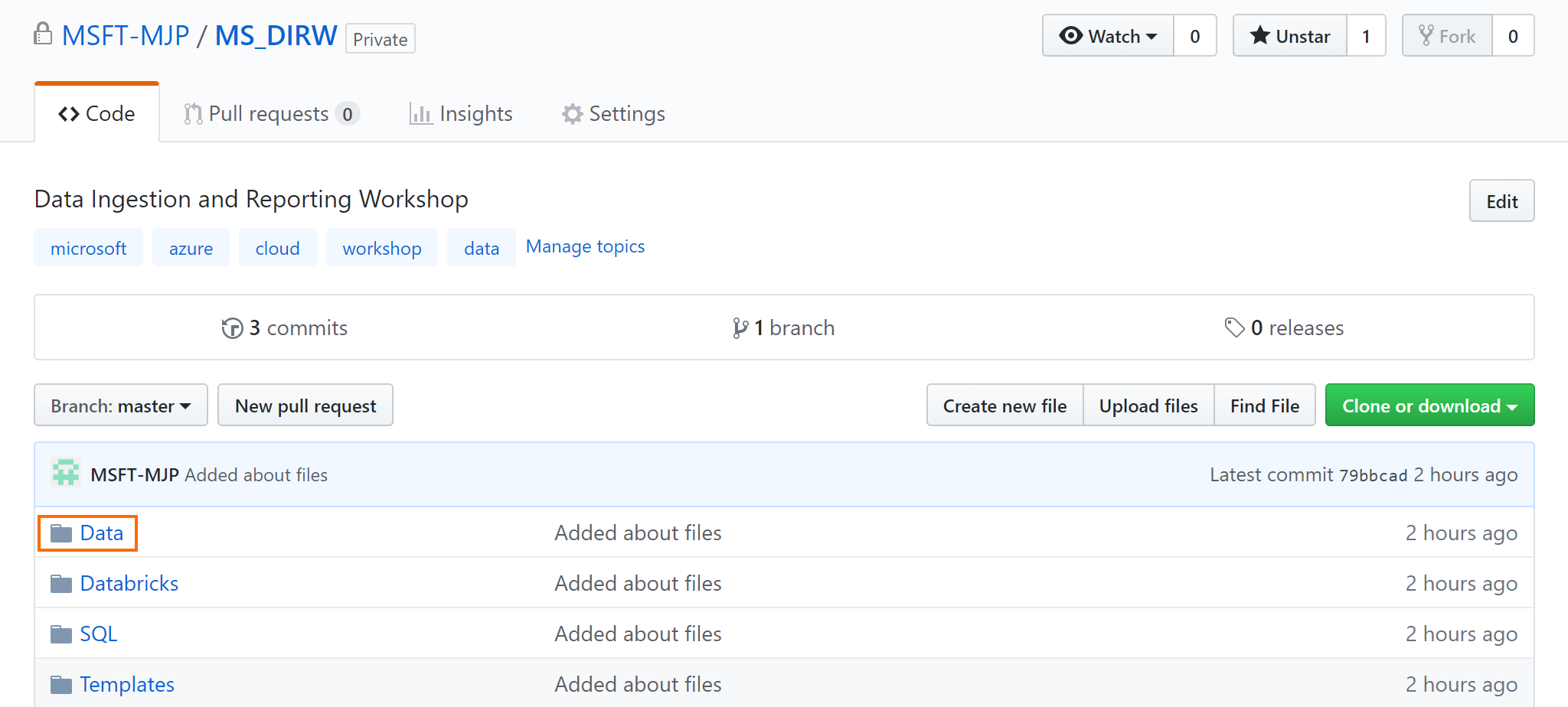


1. Create the following folder: C:\MS-DIRW

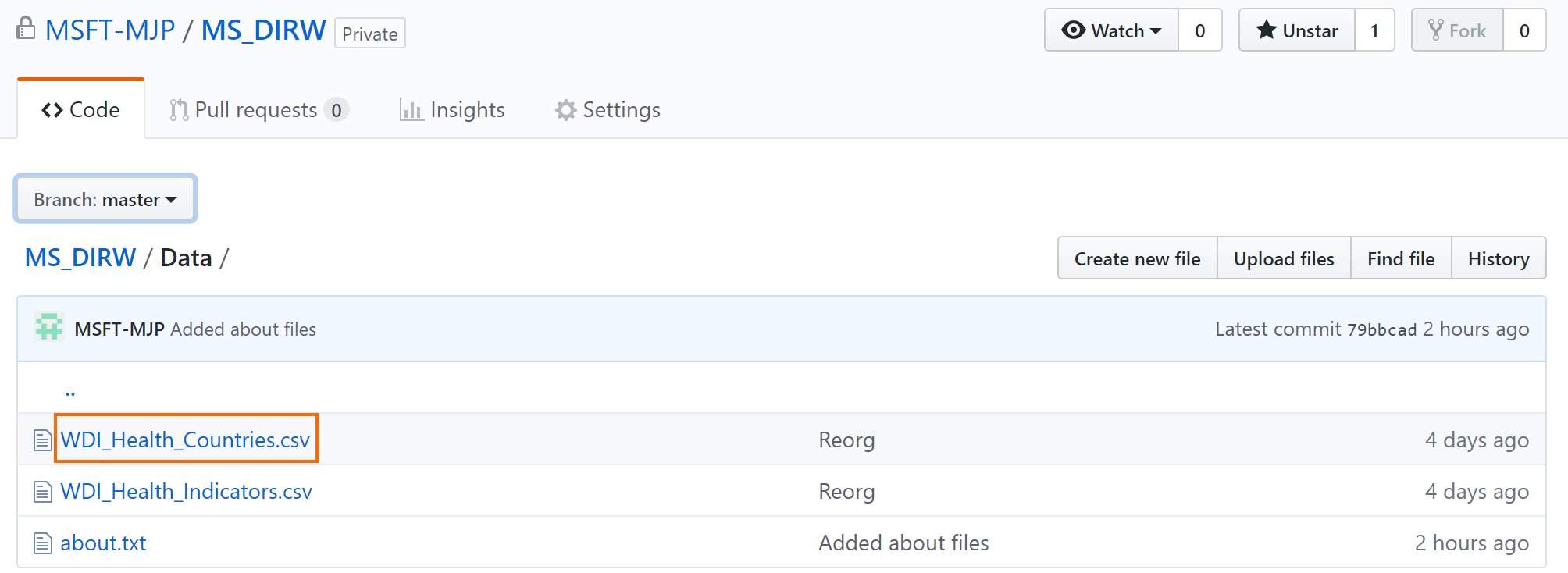


*Task 2: Download Raw Data to Local Machine*

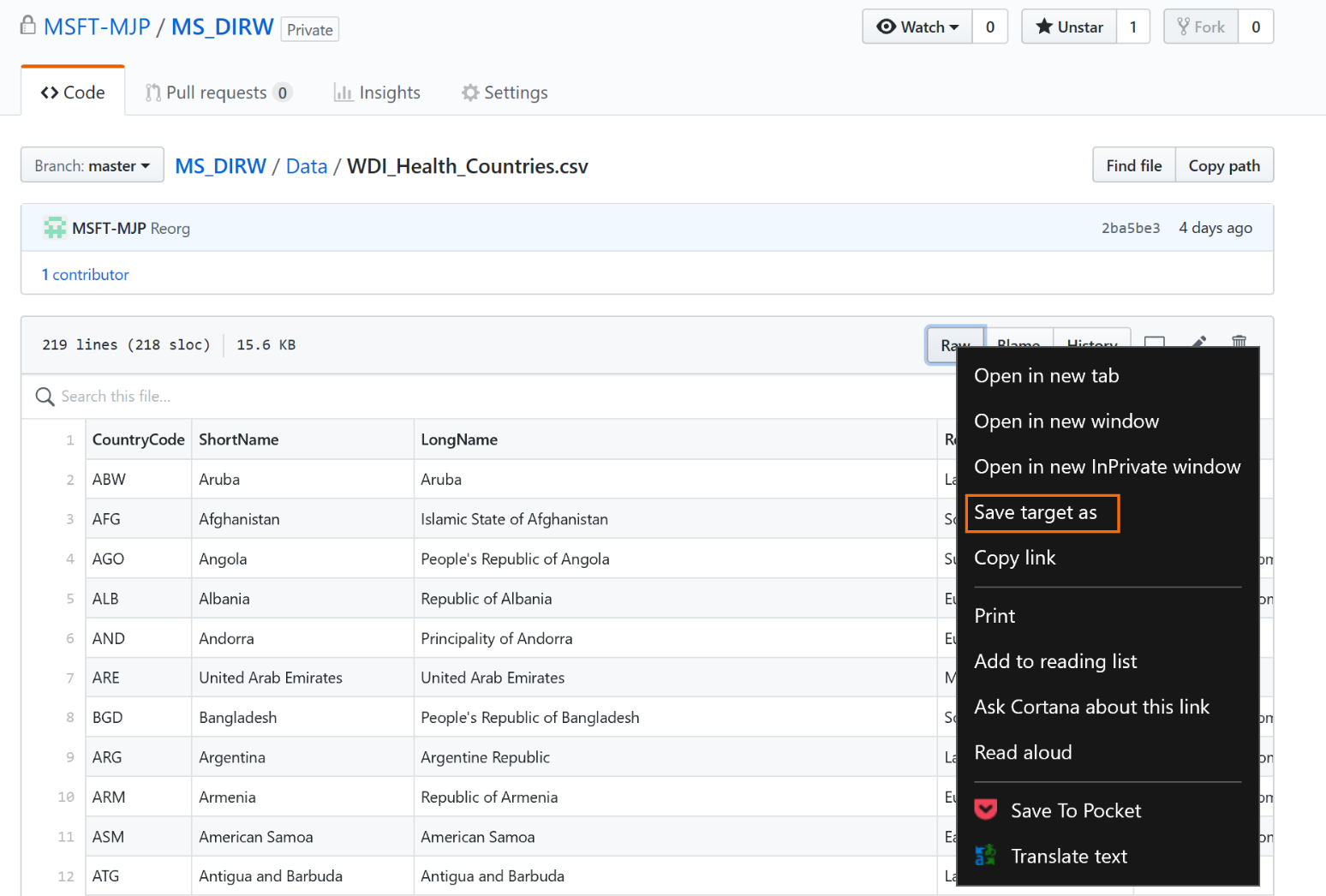
1. Navigate to the Data Folder in the GitHub MS-DIRW repository



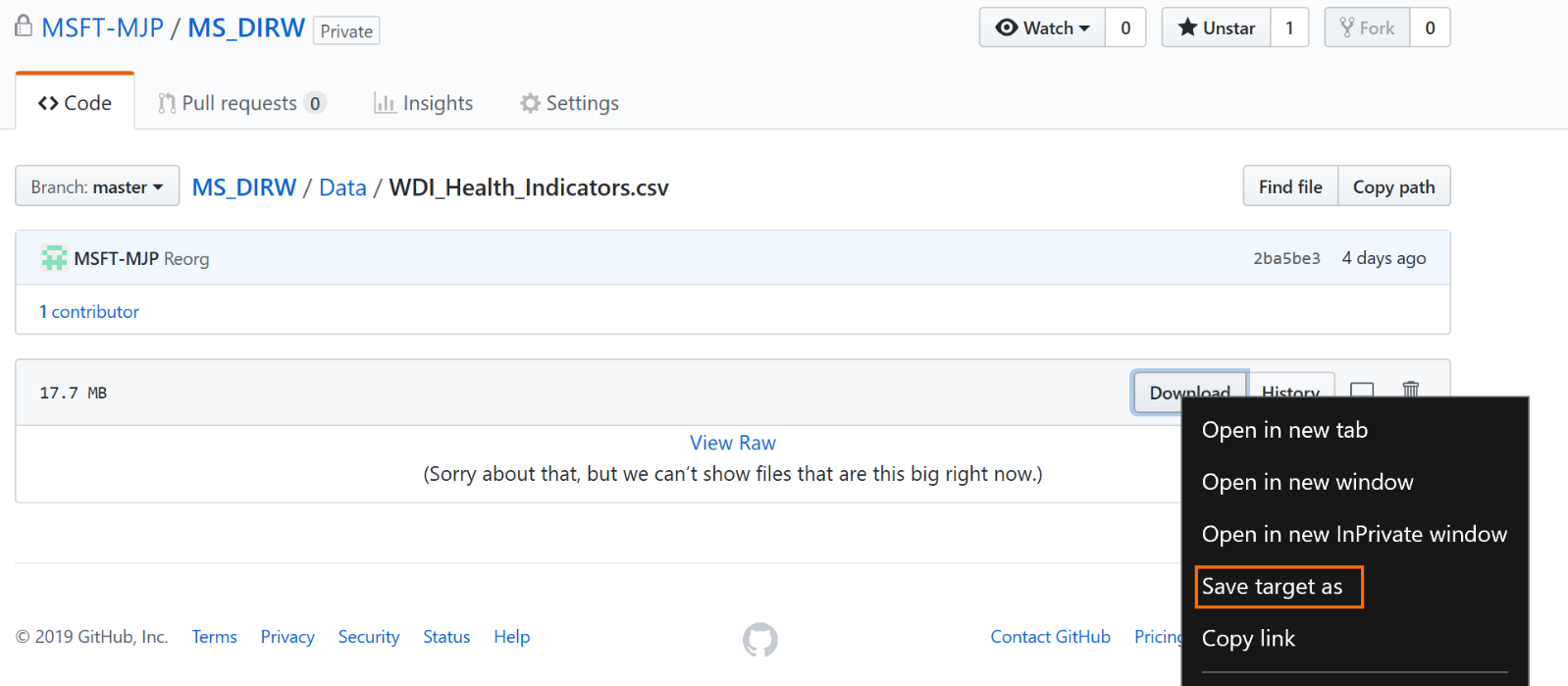
1. Click on [WDI\_Health\_Countries.csv](https://github.com/MSFT-MJP/MS_DIRW/blob/master/Data/WDI_Health_Countries.csv)



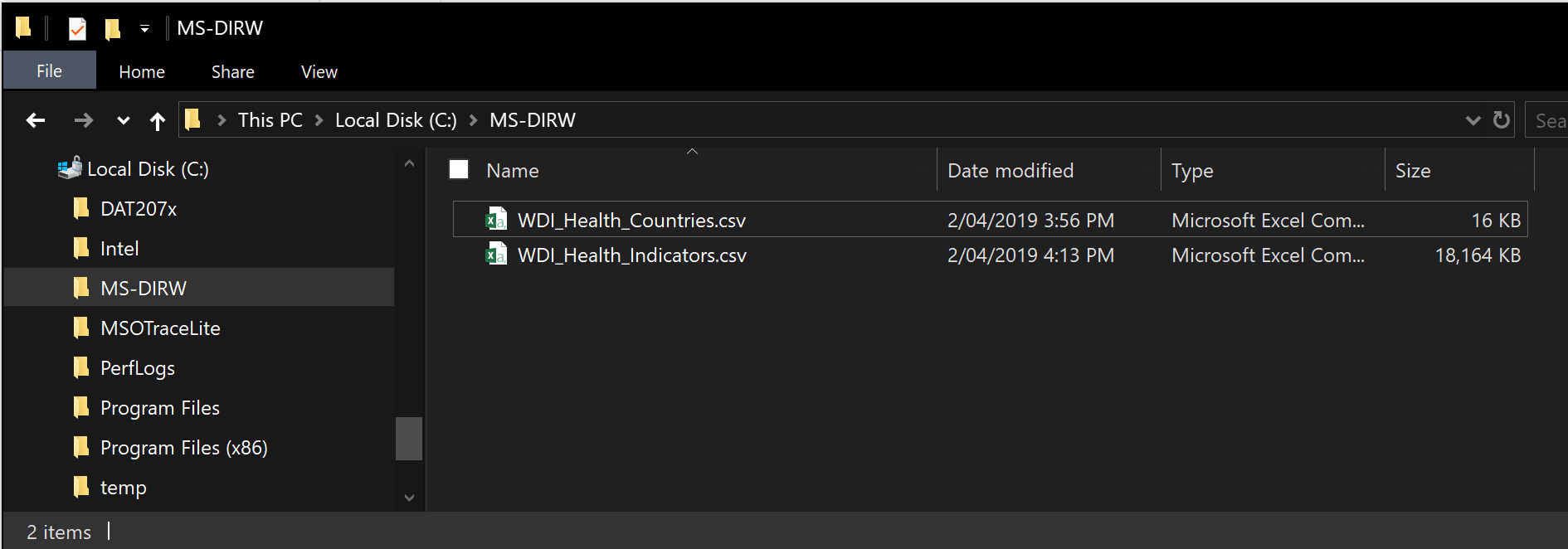
1. Right click on the Raw button and “Save target as” (or equivalent depending on your browser) and save the file in the C:\MS-DIRW folder on your local machine.



1. As the [WDI\_Health\_Indicators.csv](https://github.com/MSFT-MJP/MS_DIRW/blob/master/Data/WDI_Health_Countries.csv) file is a larger file the download option is available instead of the view Raw option. Right click on the Download button and “Save target as” (or equivalent depending on your browser) and save the file in the C:\MS-DIRW folder on your local machine.



1. Use File Explorer to ensure the following structure exists:



## Lab 02: Create Raw Data Repository in Azure

In this lab we will create the Azure data repository using Azure Blob Storage and upload our CSV files from our local repository to our Azure repository using Azure Storage Explorer.

Prerequisites

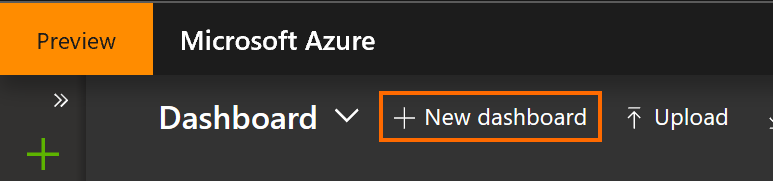
* Previous labs completed successfully
* Microsoft Azure Storage Explorer installed.

Exercise 1: Create New Dashboard

We will create a new dashboard for our workshop to make it easier to access our resources and keep our references all in one place.

*Task 1: Create a new dashboard*

1. Log into the [Azure portal](https://portal.azure.com/)
2. Click on “+ New dashboard”



1. Update the name to MS-DIRW and then click the “Done customizing” button

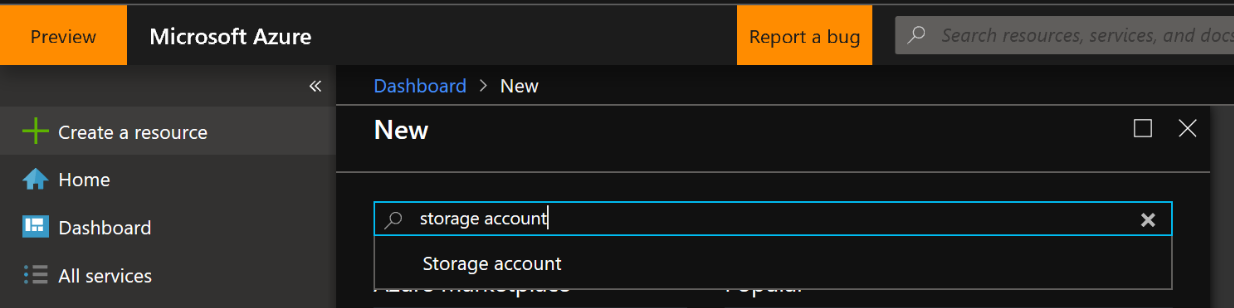


Exercise 2: Create Azure Data Repository

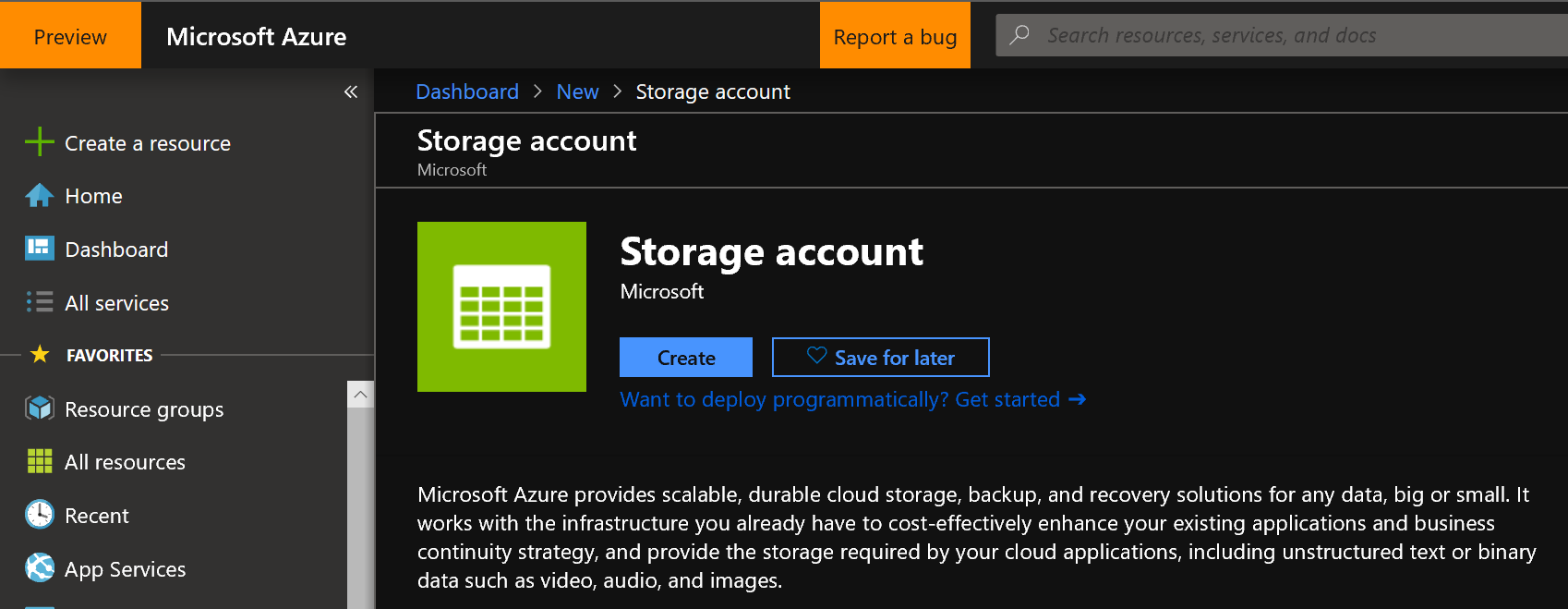
Deploy Azure Blob Storage and create the required structures.

*Task 1: Create a new Storage Account*

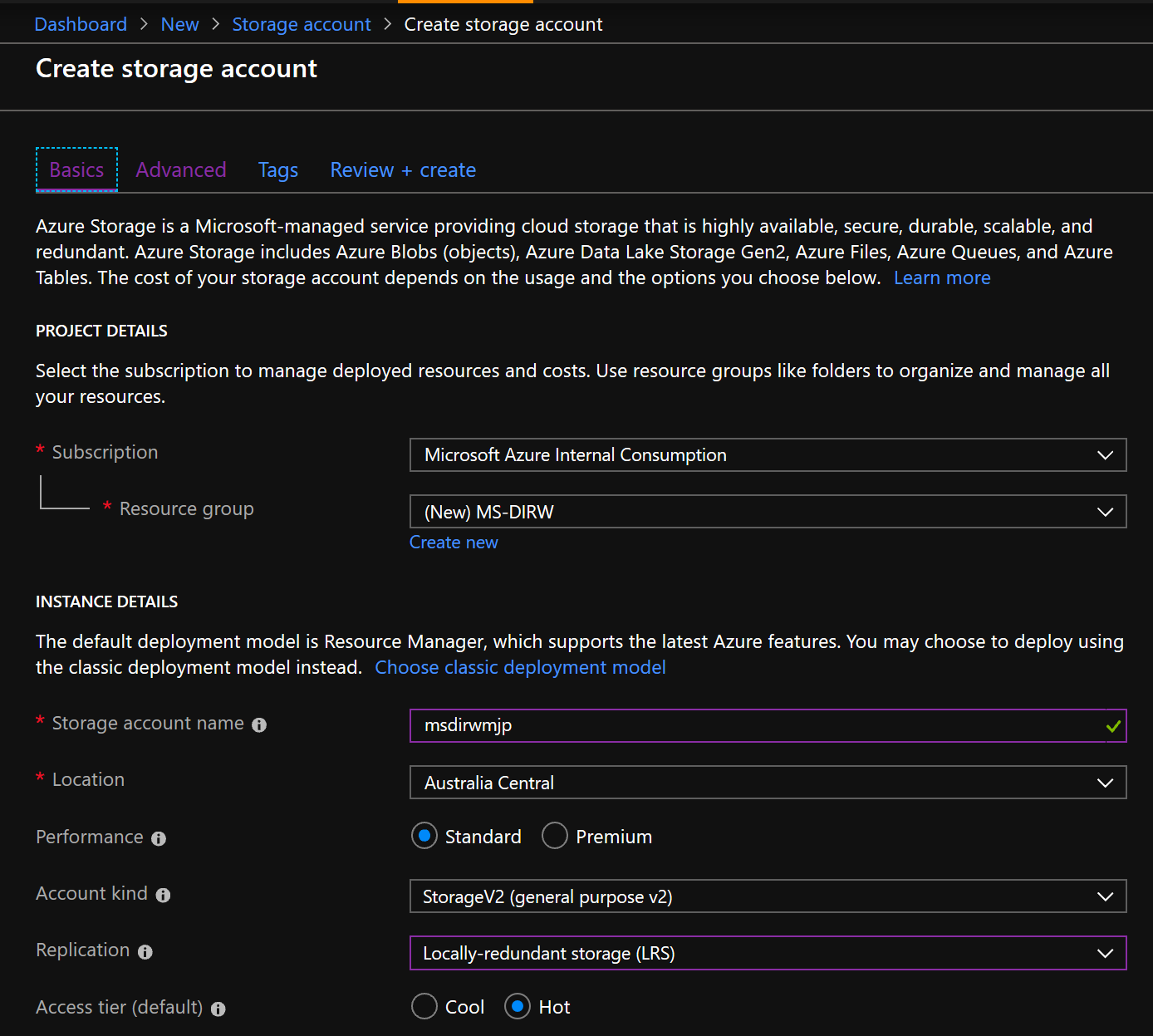
1. From your dashboard, Click on +Create a resource, type “Storage Account” and select the Storage Account entry



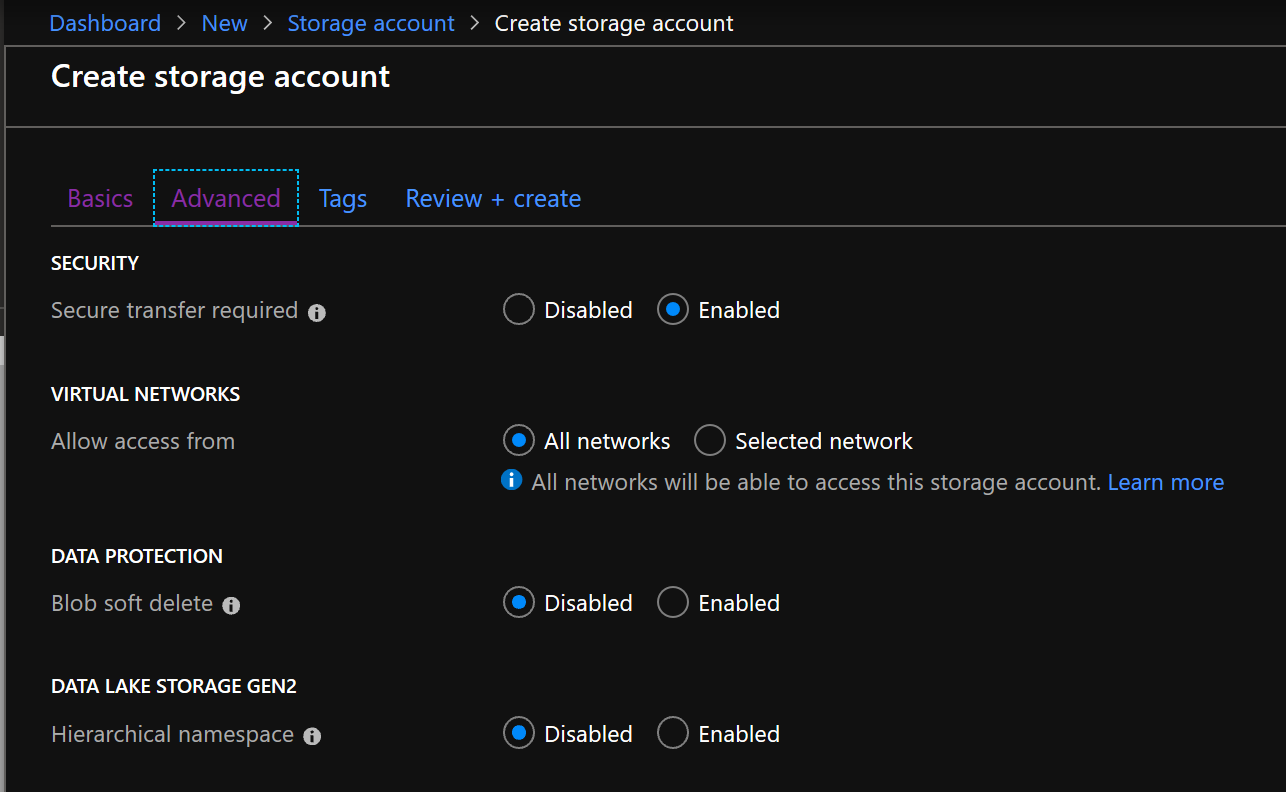
1. Click “Create”



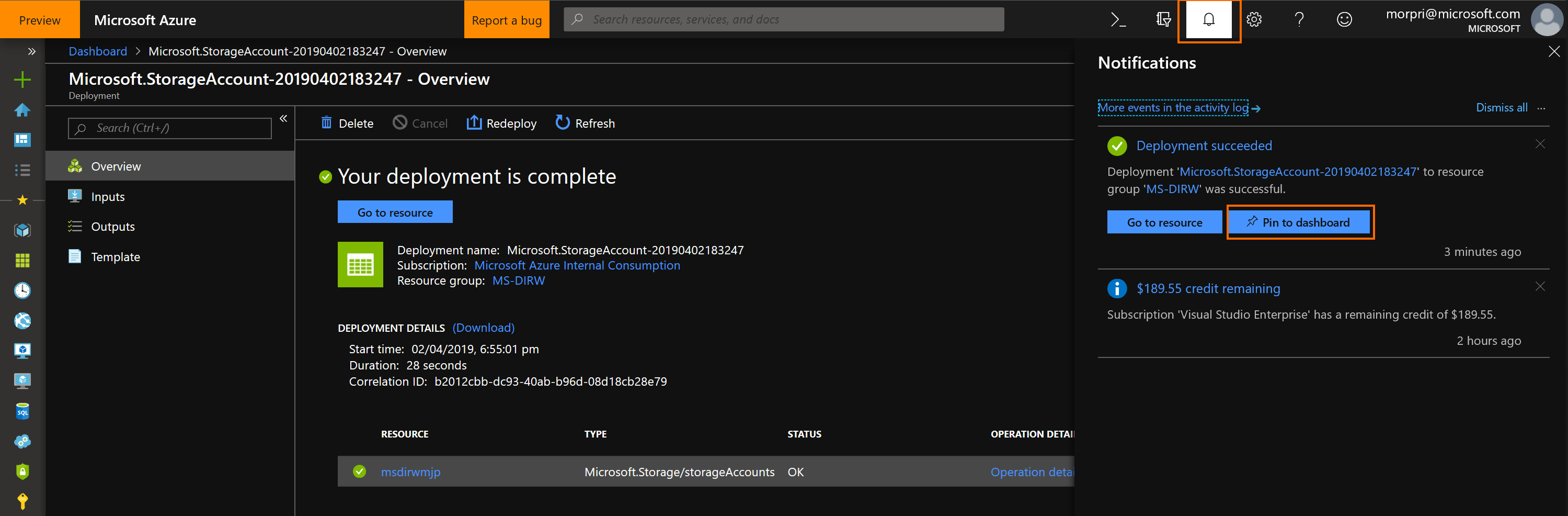
1. Fill in the following fields
   1. Subscription: Select your subscription
   2. Resource group: “Create New”, enter “MS-DIRW” (we will use this resource group for all resources created for this workshop)
   3. Storage account name: Enter “msdirw” and append your initials, this must be globally unique so add additional information if your chosen name has already been used
   4. Location: “Australia Central”, if you use a different location for performance and cost reasons use the same location for all resources created for this workshop
   5. Performance: Standard, fast enough for this workshop
   6. Account kind: StorageV2 (general purpose v2)
   7. Replication: LRS, we don’t need Geo redundant for this workshop but for production workloads Geo redundant is recommended
   8. Access tier: “Hot”



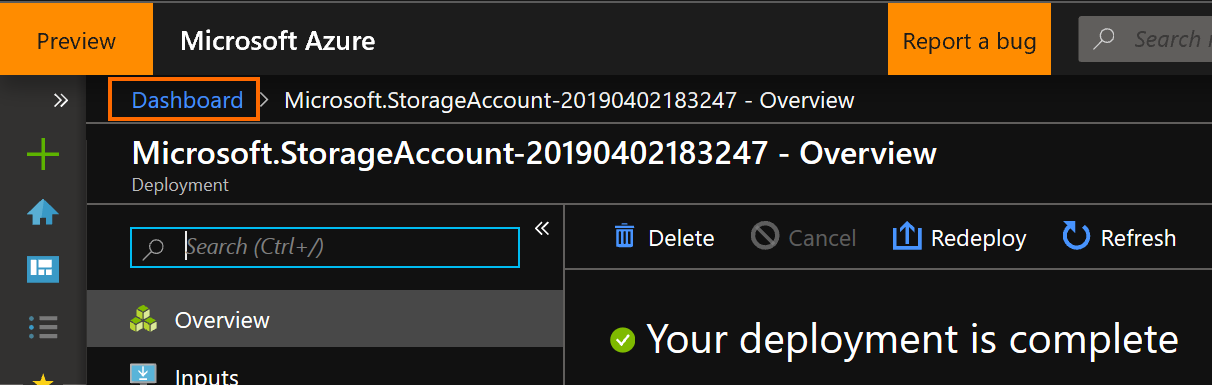
1. Under the Advanced tab we will use the defaults, note that we have enabled Secure transfer and that we can restrict access to certain networks to reduce surface area



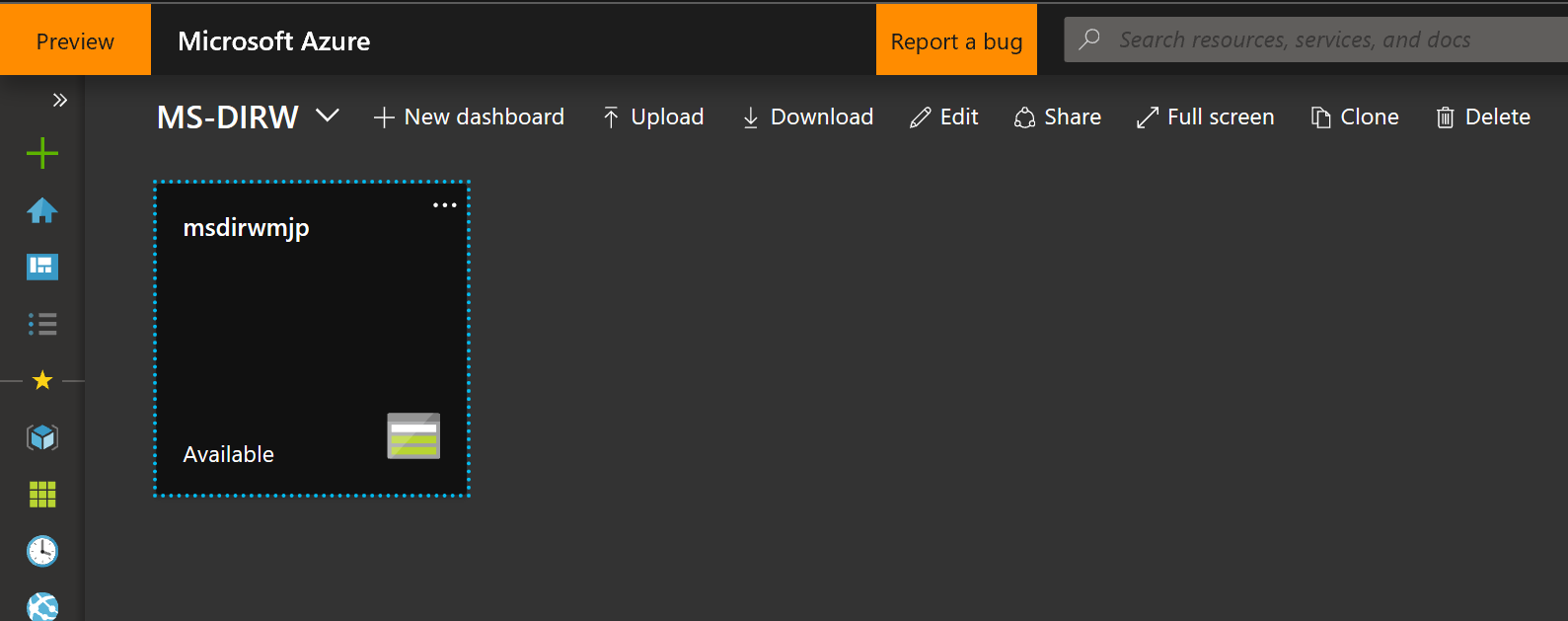
1. Review and Create your Storage account
2. When complete click on the notification (access using the little bell icon at the top right) and then click the “Pin to dashboard” button, this will make it easier to access this resource latter



1. Click on the “Dashboard” link in the top left navigation path to get back to the dashboard (you can use this at any time to “return home”)

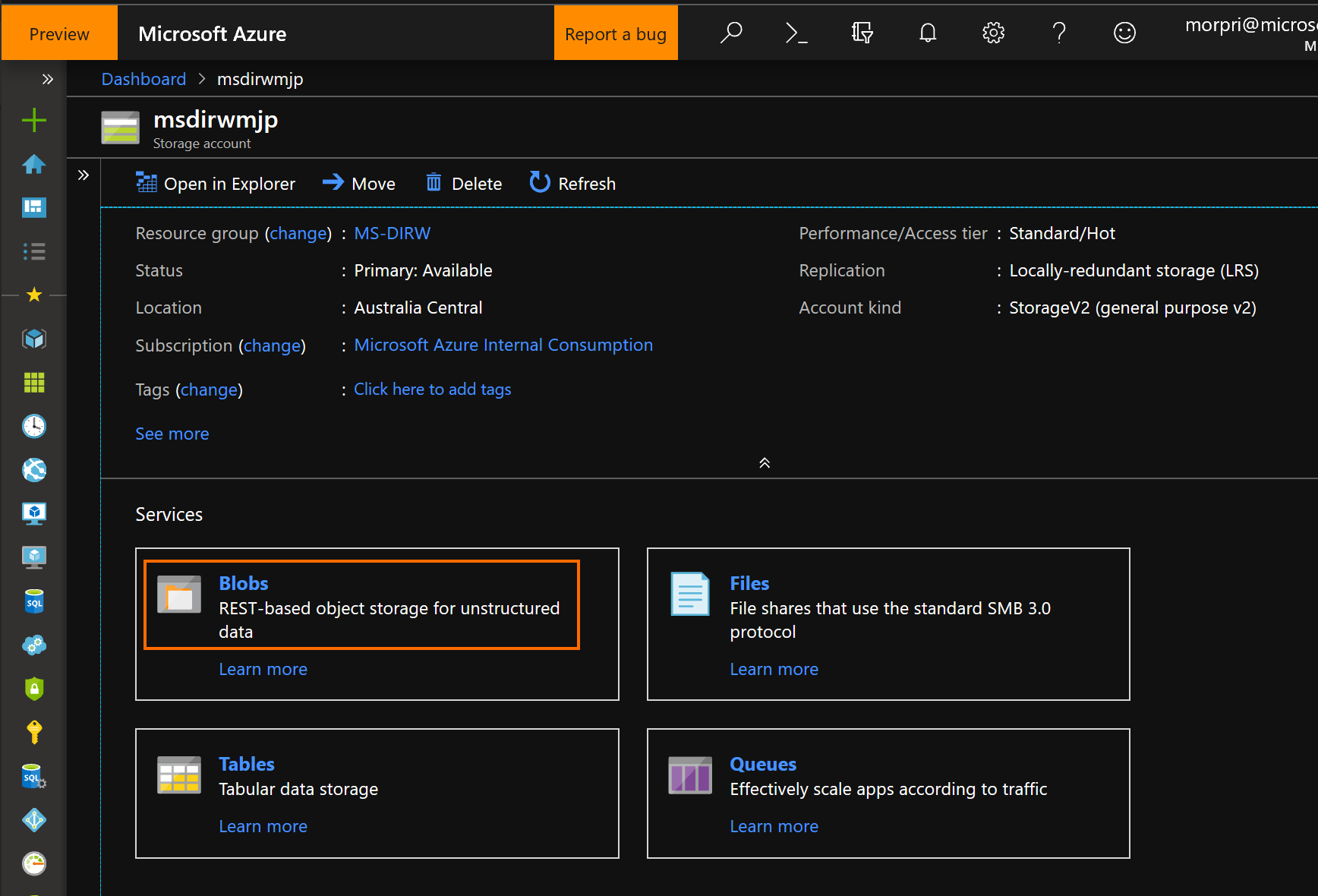


1. You should now see your new Storage account resource on your Dashboard

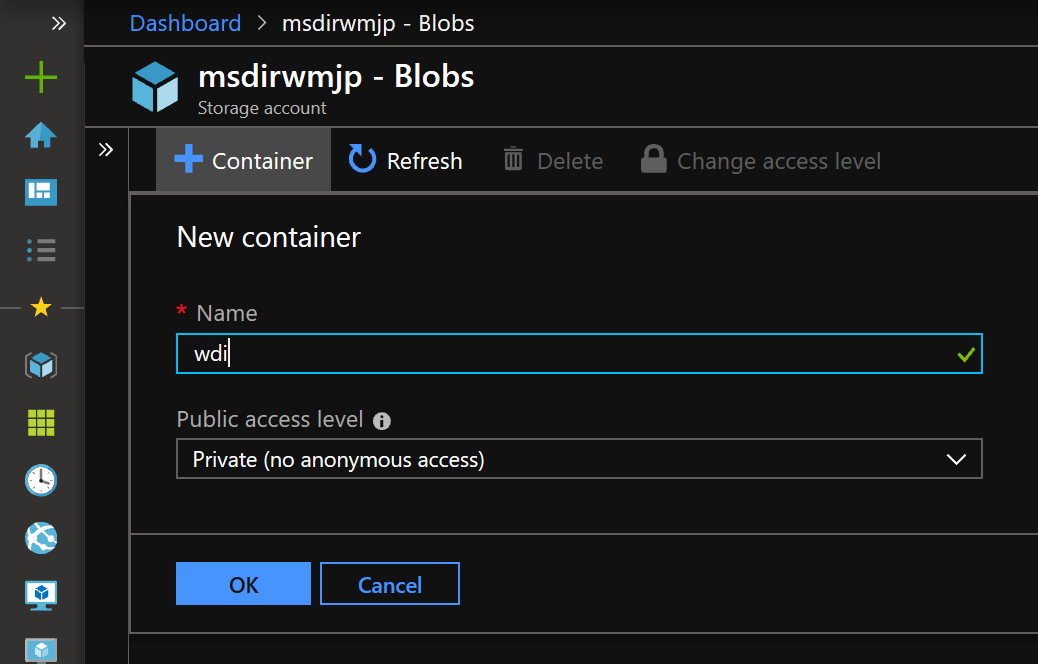


*Task 2: Create a new container to hold our raw data*

1. Click on your Storage account resource
2. Click on Blobs in the Services area



1. Click “+ Container” and name your new container “wdi”, click OK

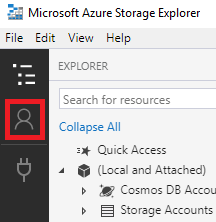


Exercise 3: Use Azure Storage Explorer (ASE) to Upload Raw Data

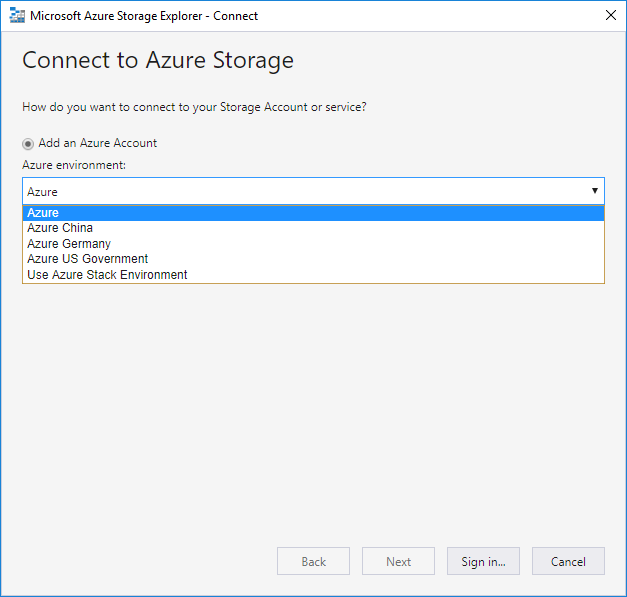
We will configure and connect ASE to our Azure data repository we have created and upload our raw data CSVs to our wdi container.

*Task 1: Connect to an Azure subscription with ASE*

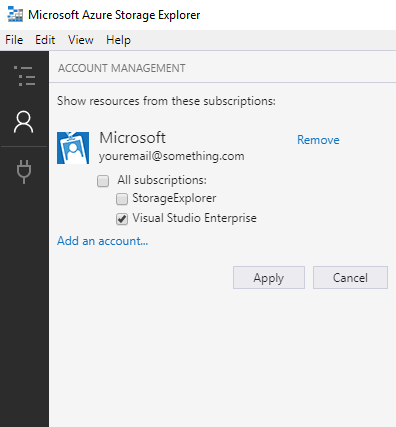
1. In Storage Explorer, select “Manage Accounts” to go to the “Account Management” panel.



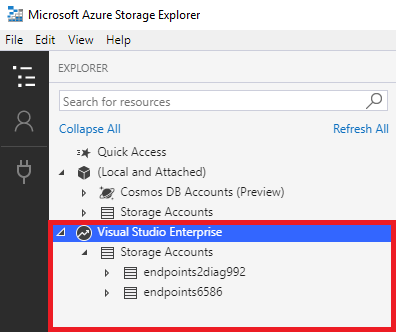
1. The left pane now displays all the Azure accounts you've signed in to. To connect to another account, select “Add an account”
2. Ensure “Azure” is selected in the environment dropdown



1. After you successfully sign in with an Azure account, the account and the Azure subscriptions associated with that account are added to the left pane. Select the Azure subscriptions that you want to work with, and then select Apply (Selecting All subscriptions: toggles selecting all or none of the listed Azure subscriptions)

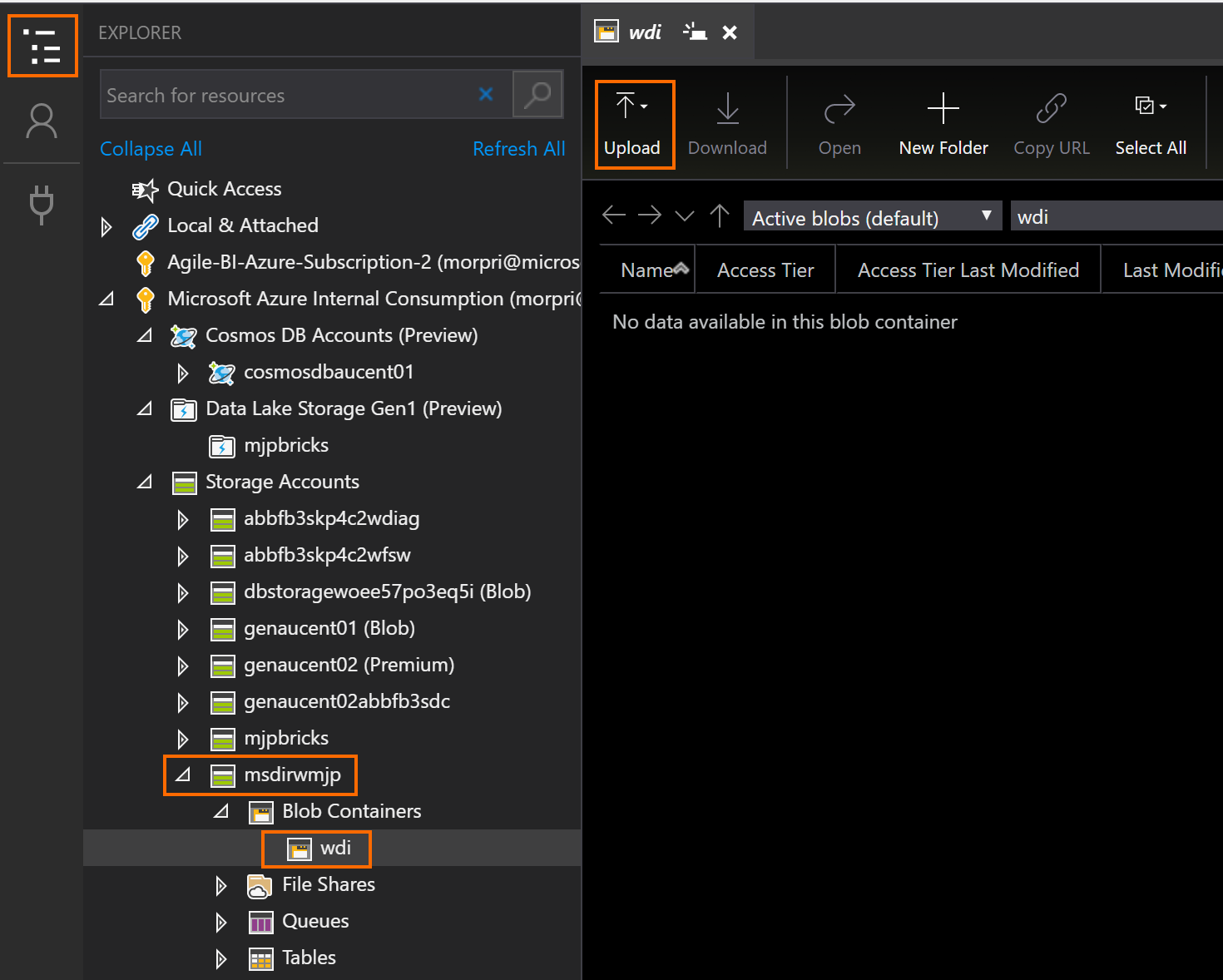


1. The left pane displays the storage accounts associated with the selected Azure subscriptions.

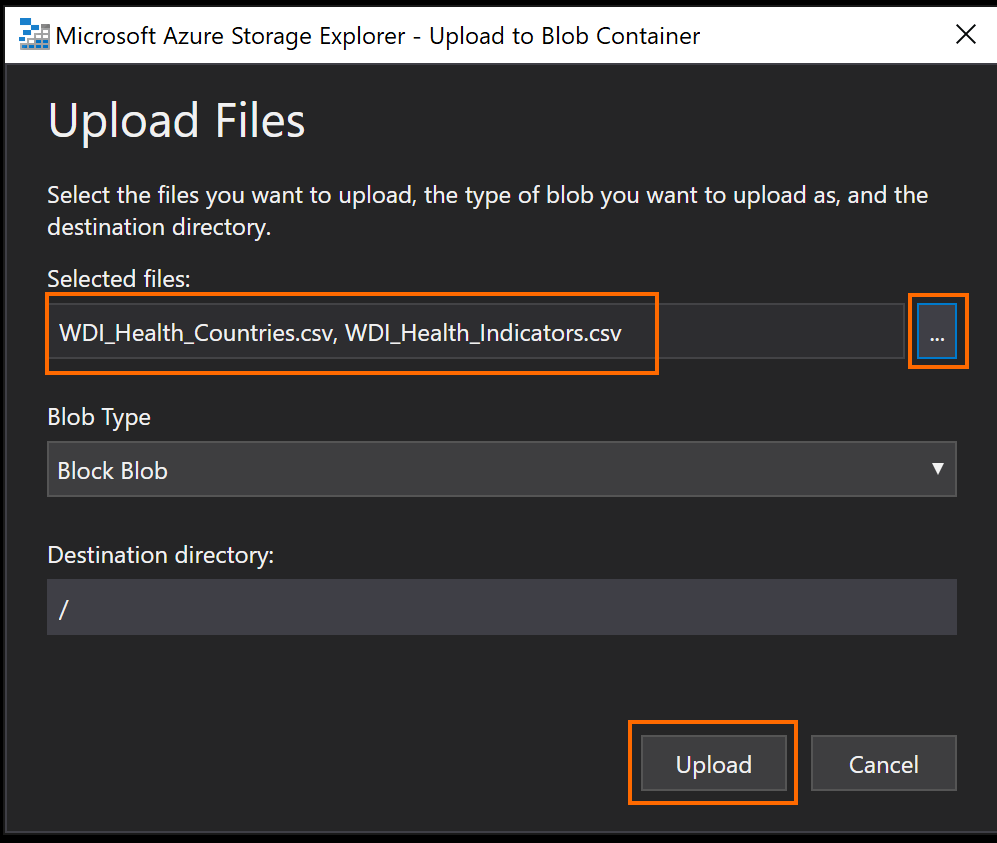


*Task 2: Upload the Raw CSV files*

1. In Storage Explorer, navigate to your wdi container you created and select “Upload”



1. Use the “…” button to navigate to C:\MS-DIRW and select our 2 raw data CSV files – WDI\_Health\_Countries.csv and WDI\_Health\_Indicators.csv. Click the Upload button



1. Refresh the view and ensure it looks like this:

