Microsoft Cloud Workshops

Data Ingestion and Reporting

Walkthrough

April 2019

© 2019 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](#_Toc5182782)

[Objectives 1](#_Toc5182783)

[The Workshop 1](#_Toc5182784)

[Delivery 1](#_Toc5182785)

[Duration 2](#_Toc5182786)

[Requirements 2](#_Toc5182787)

[Workshop Step-by-Step 3](#_Toc5182788)

[Lab 01: Setup 3](#_Toc5182789)

[Exercise 1: Access Azure 3](#_Toc5182790)

[Exercise 2: Access GitHub Repository 4](#_Toc5182791)

[Exercise 3: Create Local File Storage Location 5](#_Toc5182792)

[Lab 02: Create Raw Data Repository in Azure 9](#_Toc5182793)

[Exercise 1: Create New Dashboard 9](#_Toc5182794)

[Exercise 2: Create Azure Data Repository 9](#_Toc5182795)

[Exercise 3: Use Azure Storage Explorer (ASE) to Upload Raw Data 15](#_Toc5182796)

# 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.

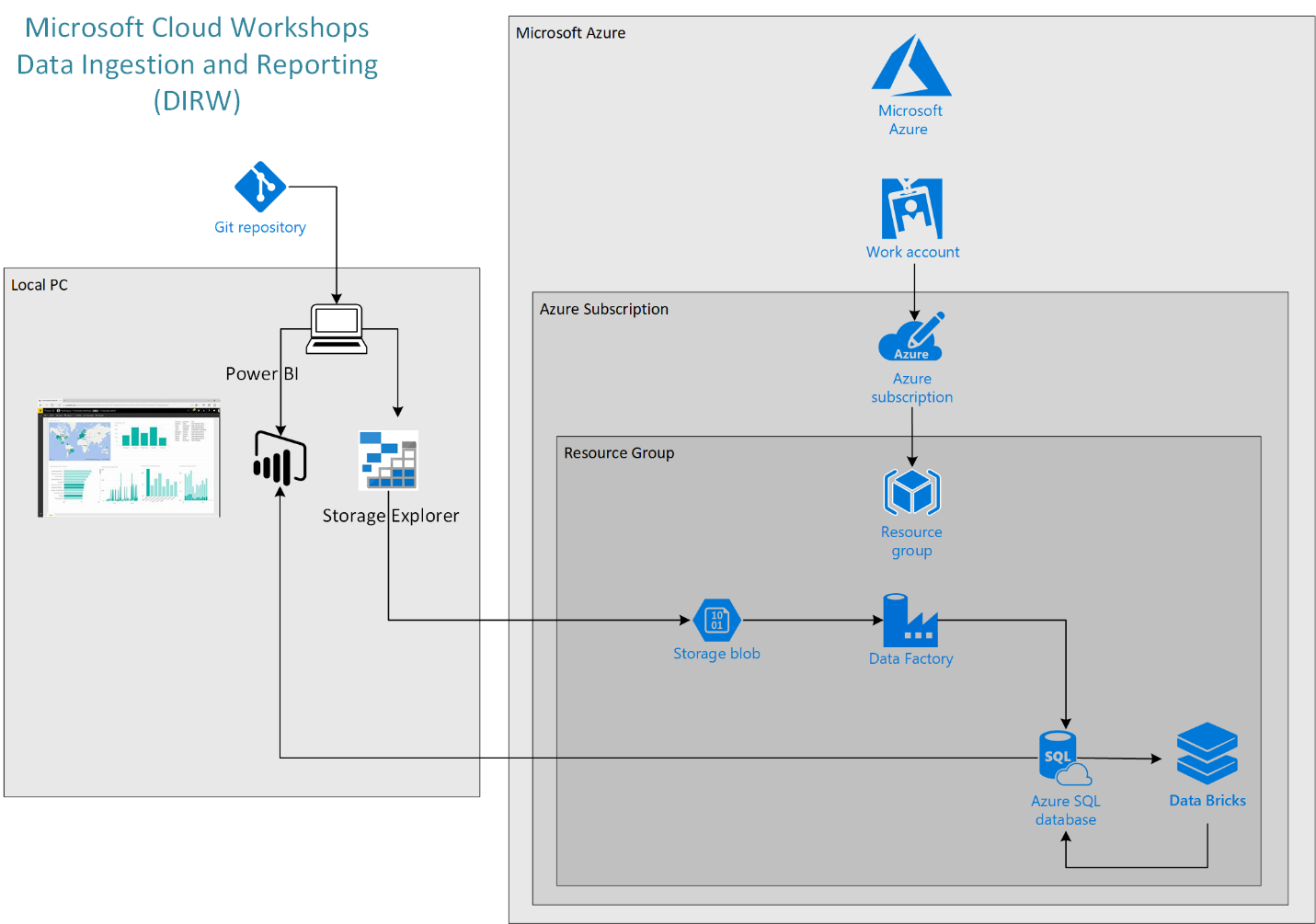
## Duration

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/)>
  + Azure Data Studio installed (latest version: <https://docs.microsoft.com/en-us/sql/azure-data-studio/download?view=sql-server-2017>)

# 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. In addition we will download required scripts also.

Prerequisites

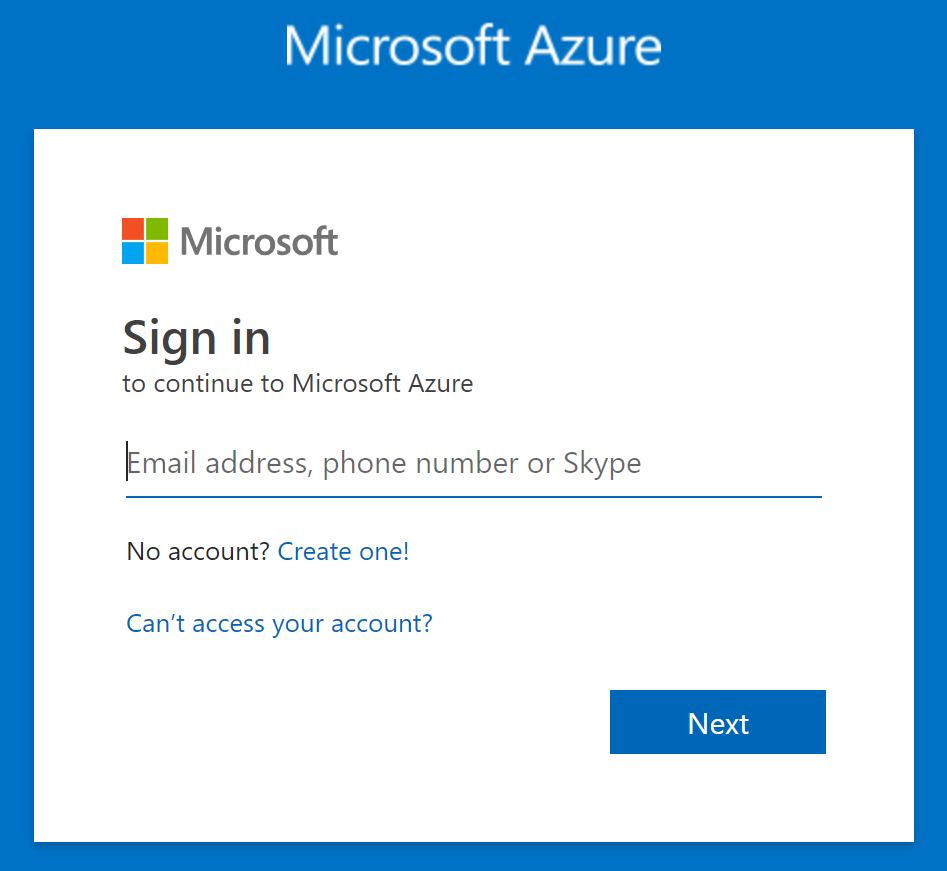
* Microsoft Azure subscription, please have your subscription details available for this workshop. If you do not have your own subscription (recommended), a free trial subscription can be accessed here: <http://azure.microsoft.com/en-us/pricing/free-trial/>

### 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:



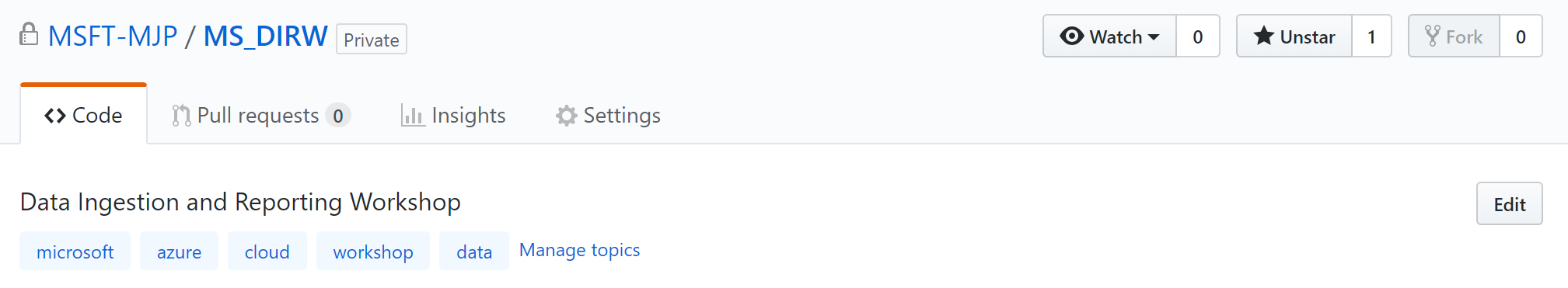
1. You should now see your Azure dashboard, bookmark this page for future reference

### 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>



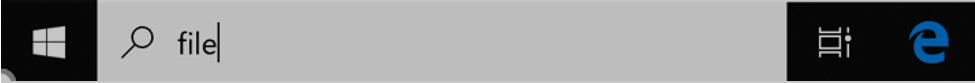
1. Bookmark this repository for future reference

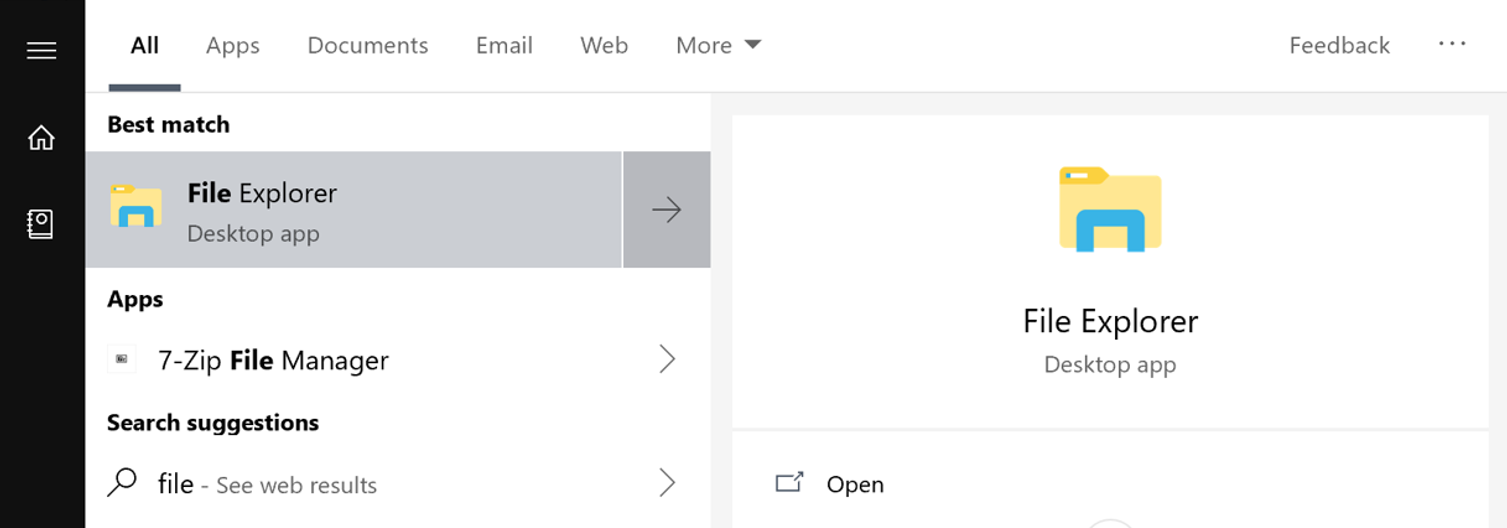
### Exercise 3: Create Local File Storage Location

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

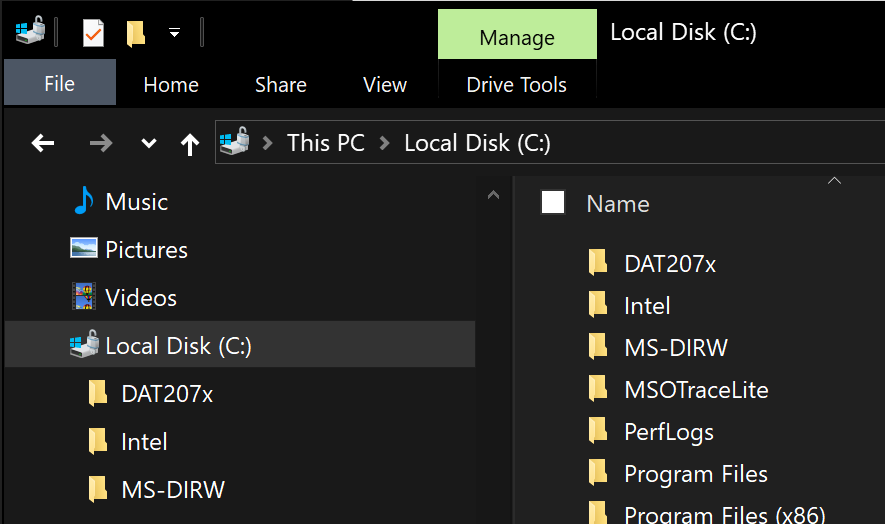
*Task 1: Create local folder*

1. Open File Explorer



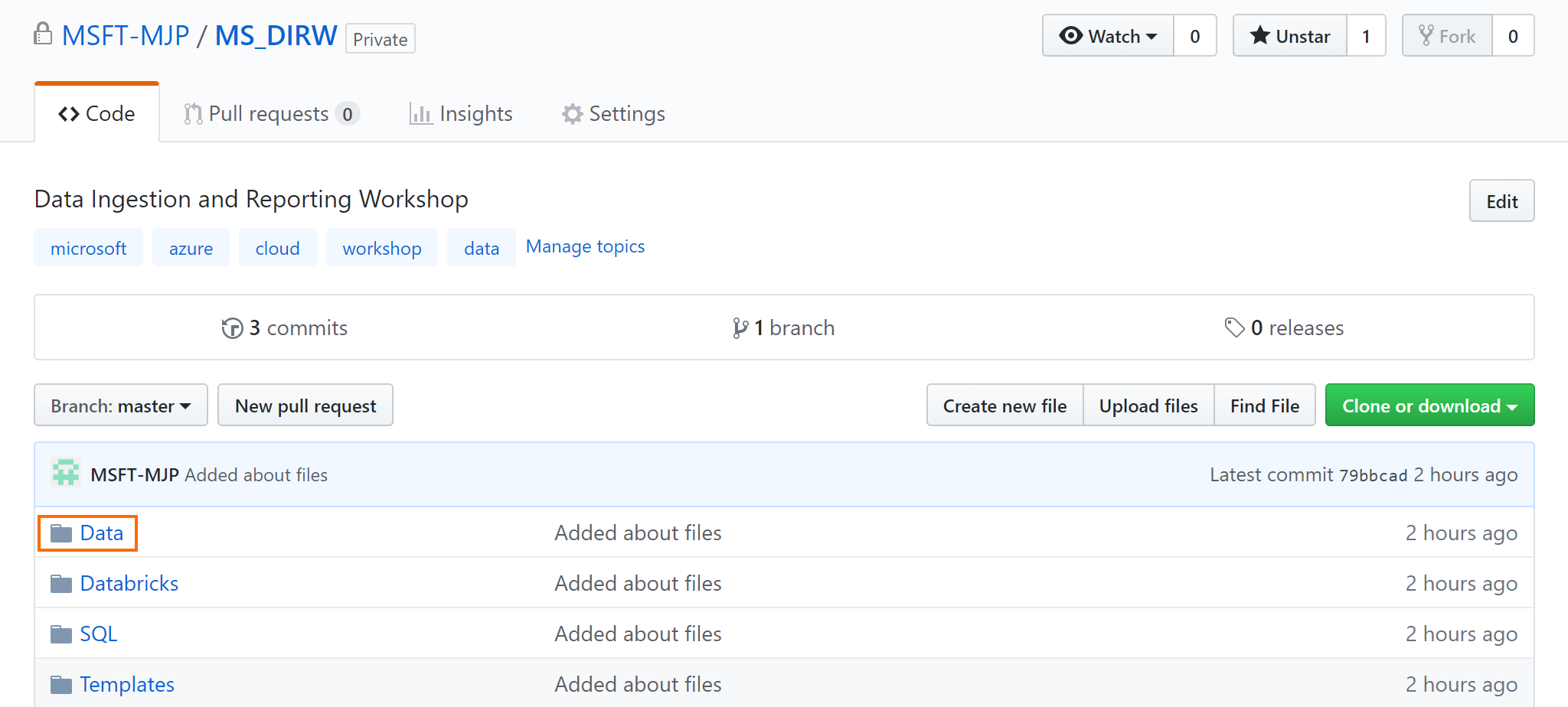


1. Create the following folder: “C:\MS-DIRW”

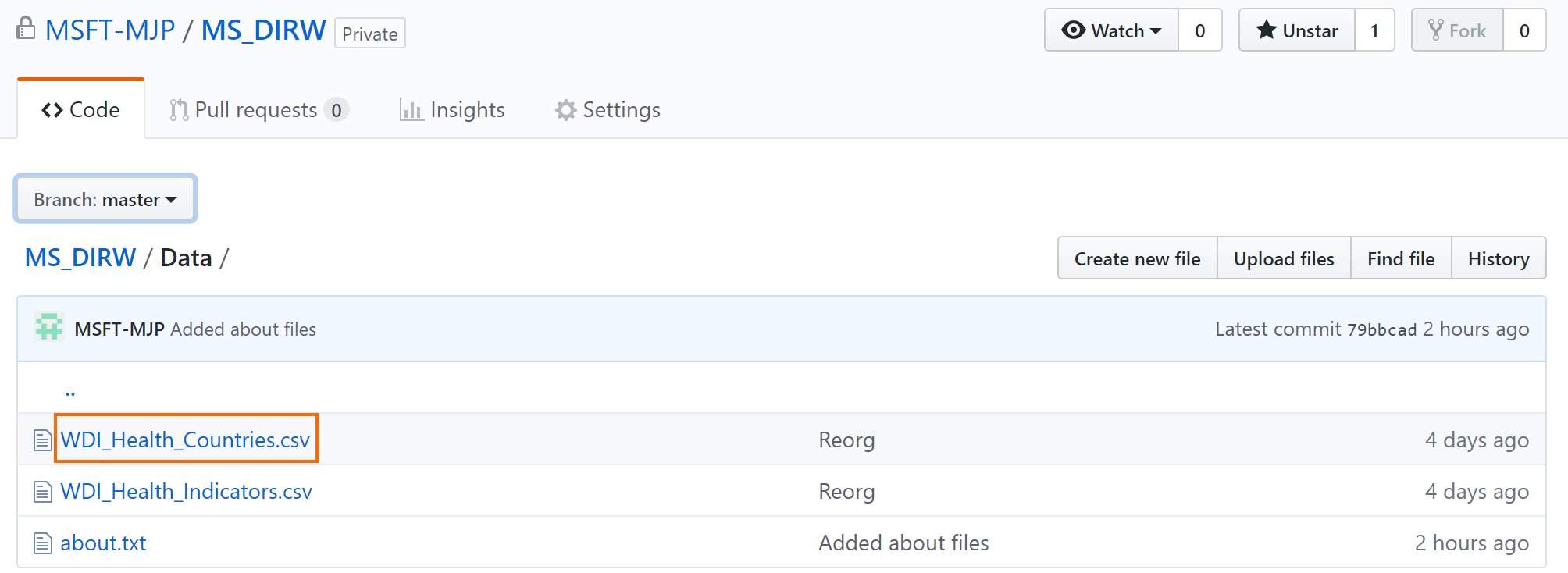


*Task 2: Download Raw Data to the Local Machine*

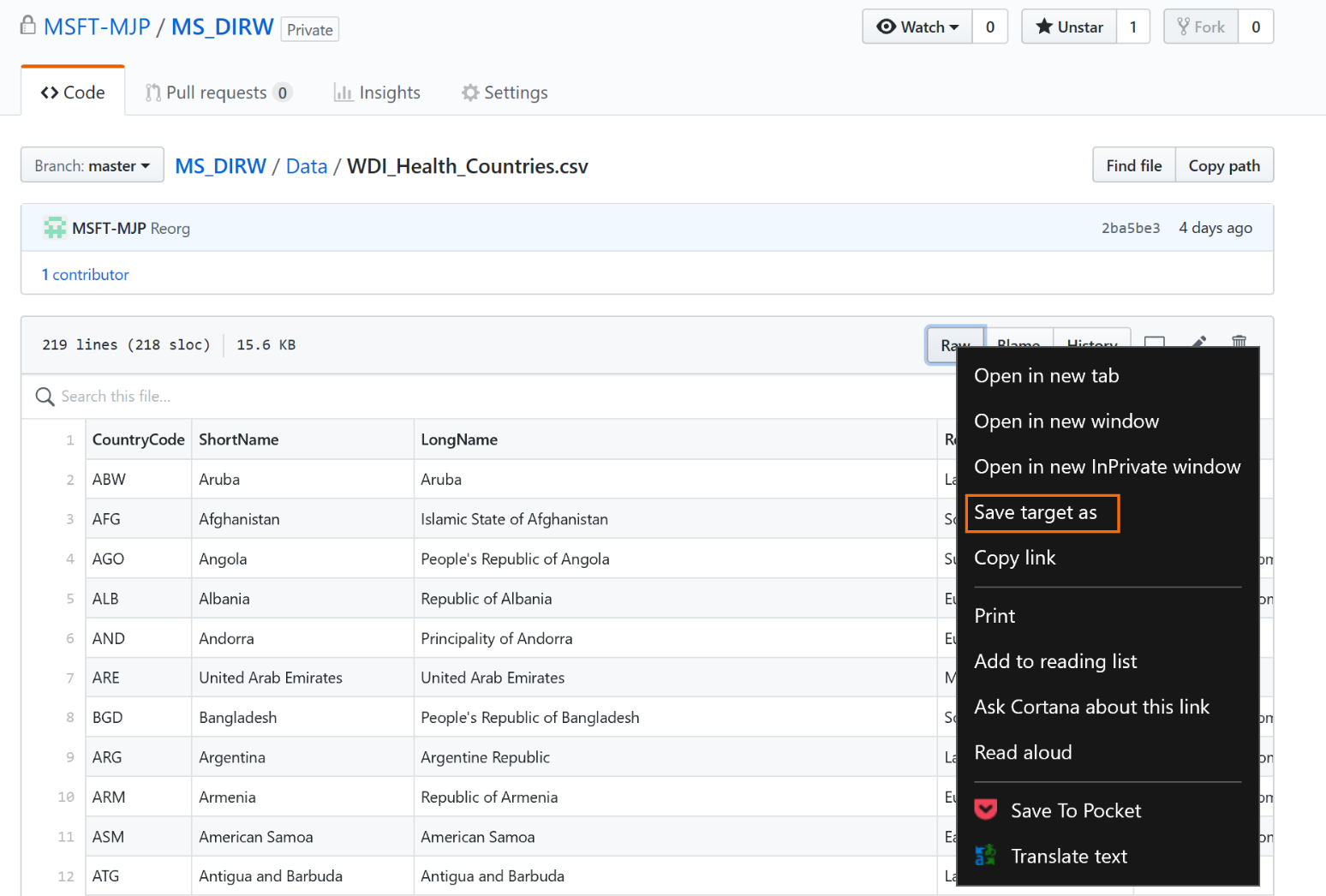
1. Navigate to the [Data Folder](https://github.com/MSFT-MJP/MS_DIRW/tree/master/Data) in the GitHub MS-DIRW repository



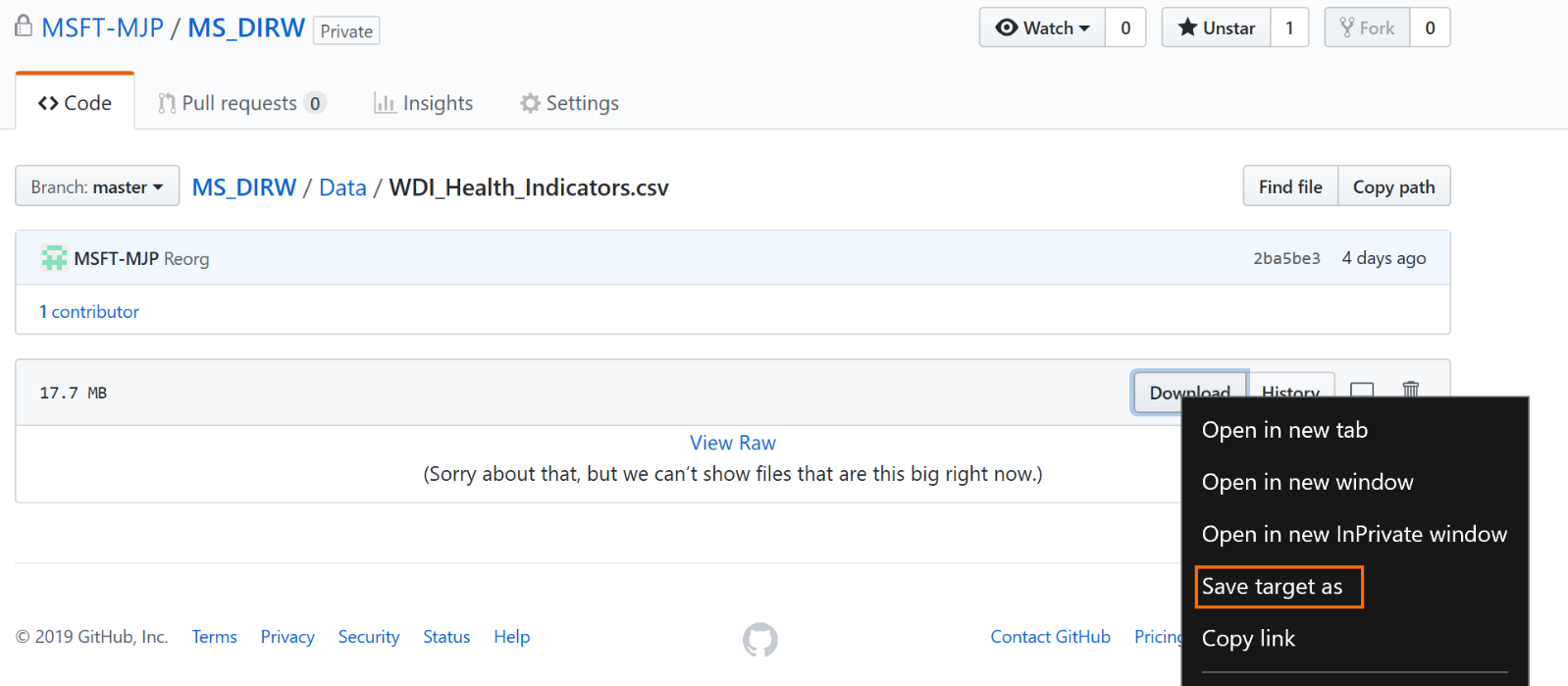
1. Click on [WDI\_Health\_Countries.csv](https://github.com/MSFT-MJP/MS_DIRW/blob/master/Data/WDI_Health_Countries.csv)



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.

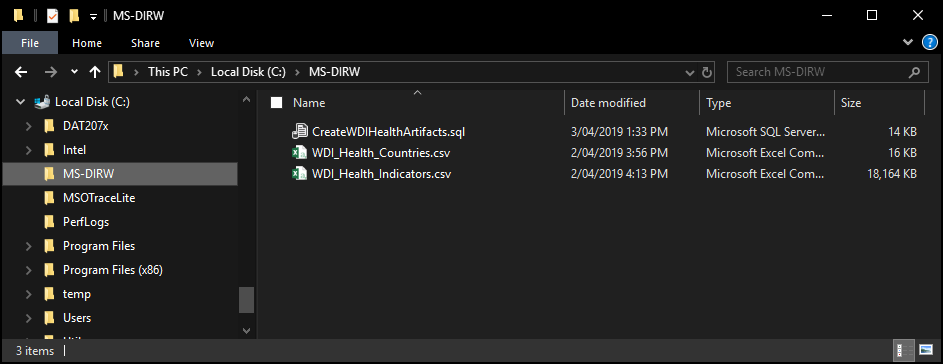


1. As the [WDI\_Health\_Indicators.csv](https://github.com/MSFT-MJP/MS_DIRW/blob/master/Data/WDI_Health_Countries.csv) file is a larger file the download option is available instead of the view Raw option. Right click on the “Download” button and “Save target as” (or equivalent depending on your browser)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



*Task 3: Download SQL notebook to the Local Machine*

1. Navigate to the [SQL Folder](https://github.com/MSFT-MJP/MS_DIRW/tree/master/SQL) in the GitHub MS-DIRW repository
2. Click on [CreateWDIHealthArtifacts.sql](https://github.com/MSFT-MJP/MS_DIRW/blob/master/SQL/CreateWDIHealthArtifacts.sql)
3. 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.
4. 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

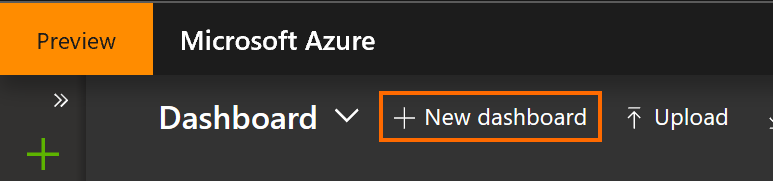
* Previous labs completed successfully
* Microsoft Azure Storage Explorer installed.

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

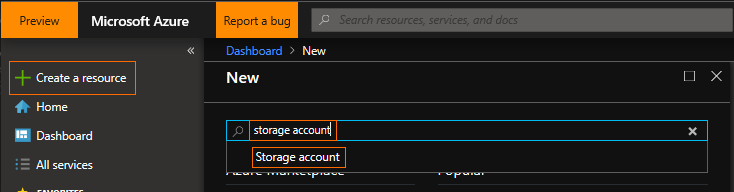


Exercise 2: Create Azure Data Repository

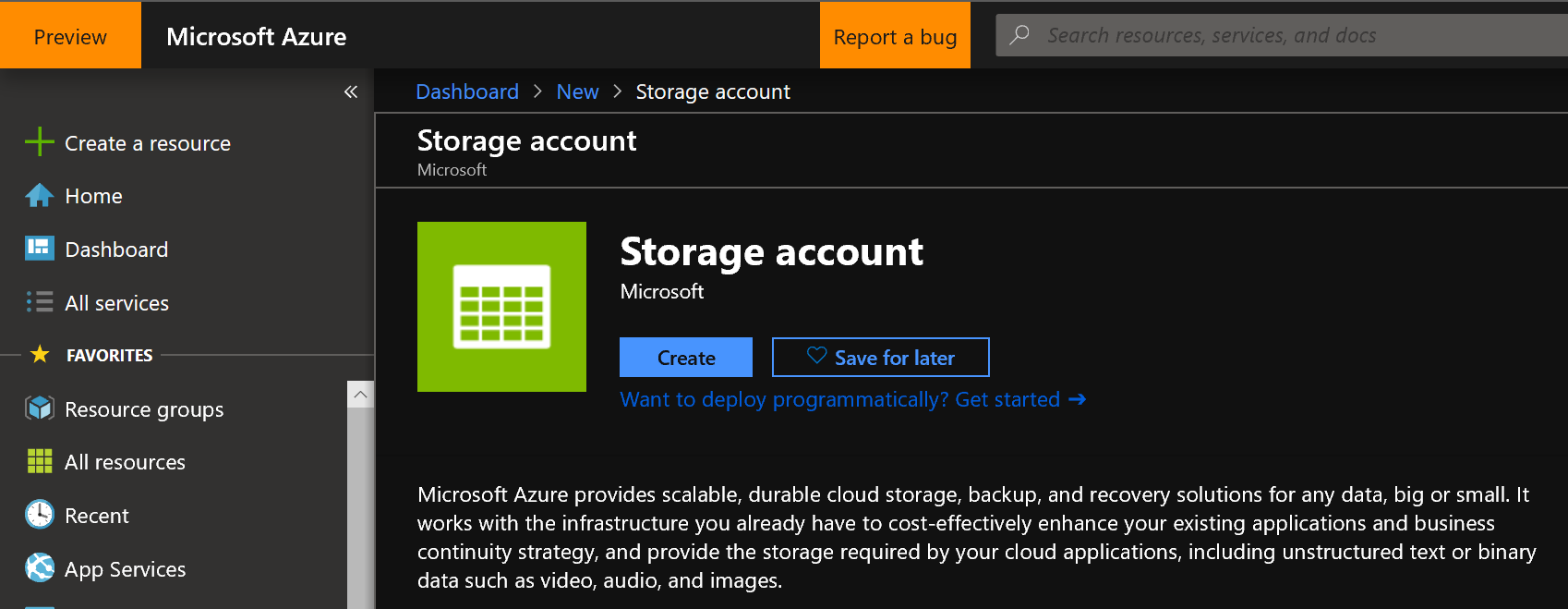
Deploy Azure Blob Storage and create the required structures.

*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, e.g. “msdirmjp”. This name must be globally unique so add additional information if your chosen name has already been used

Remember the name you have chosen so you can reuse this resource group throughout the workshop.

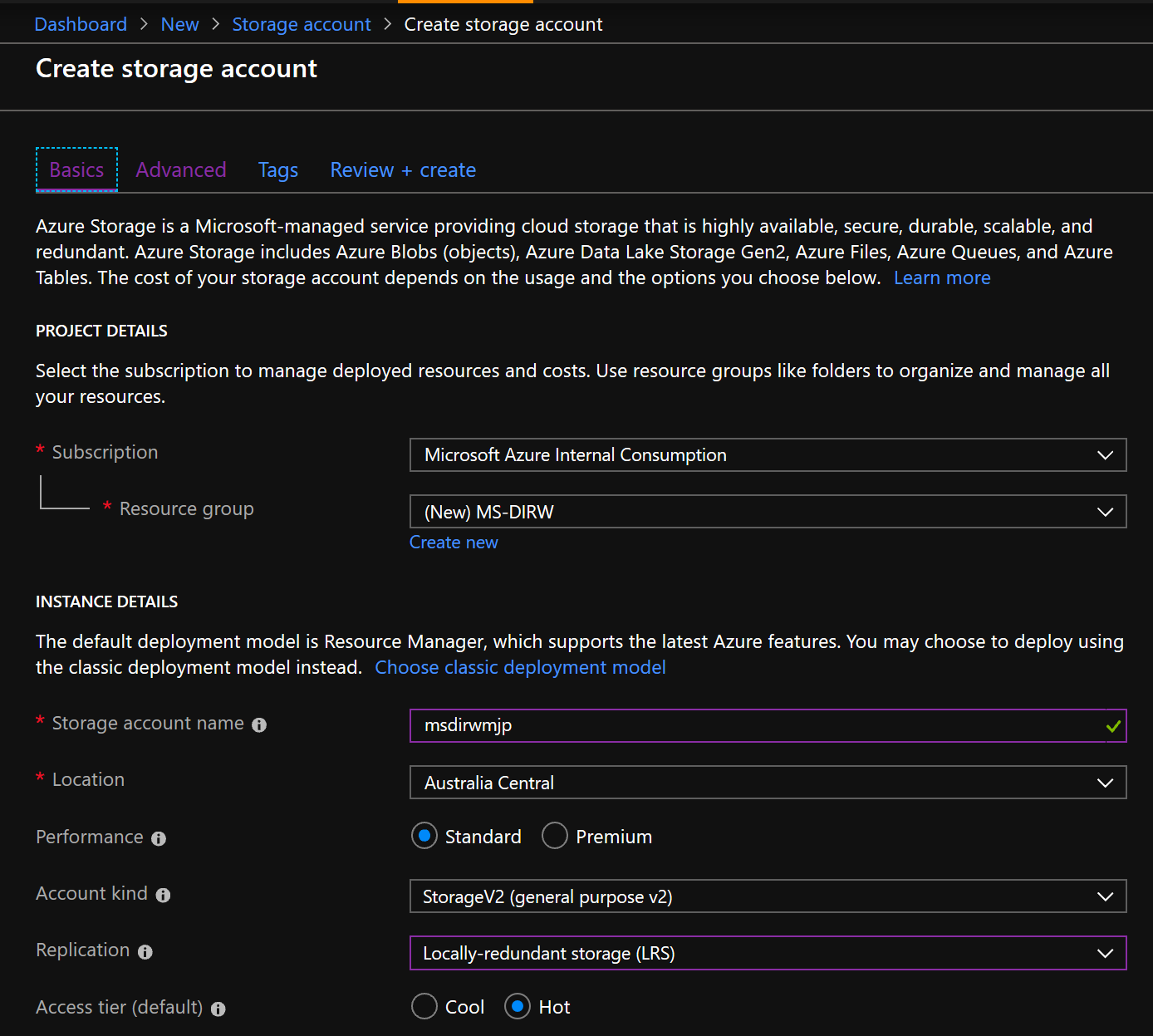
* 1. Location: “Australia Central”

If you use a different location, for performance and cost reasons, use the same location for all resources created for this workshop.

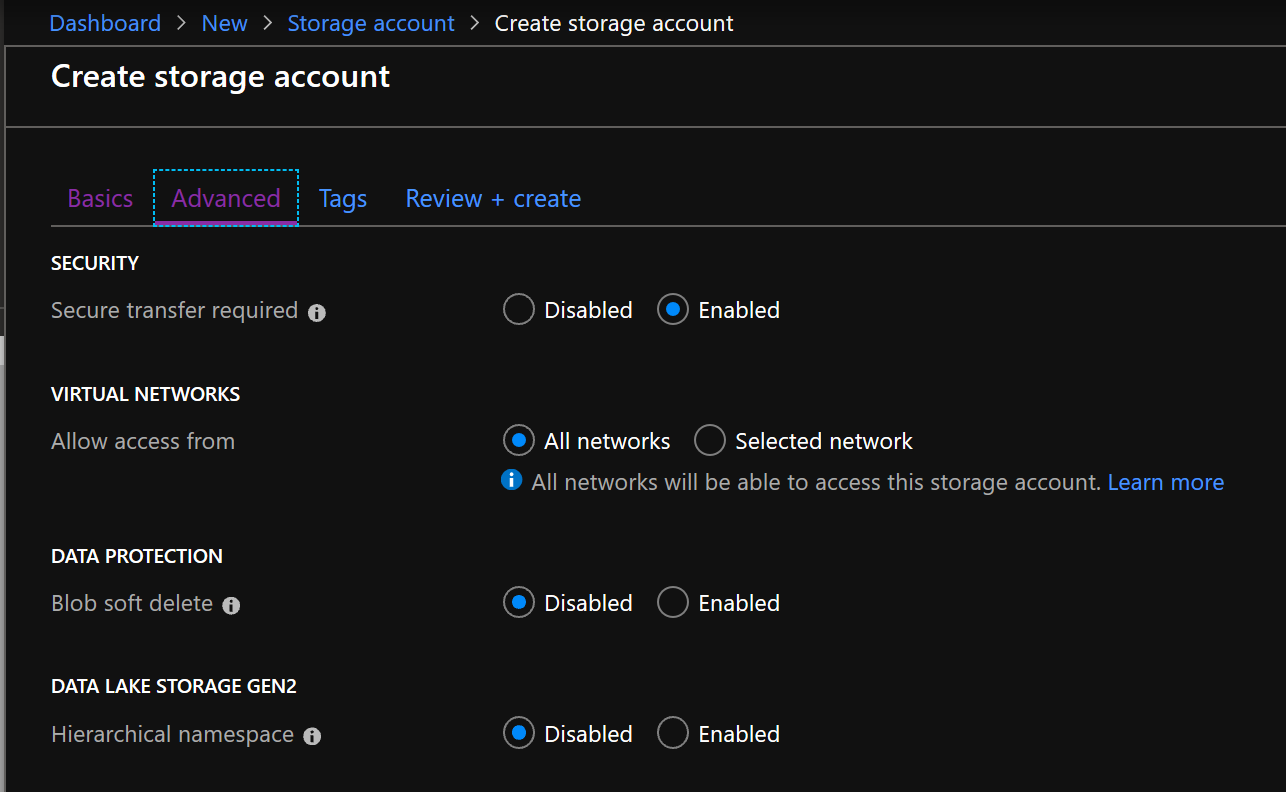
* 1. Performance: Standard, fast enough for this workshop
  2. Account kind: StorageV2 (general purpose v2)
  3. Replication: LRS

We don’t need geo-redundancy for this workshop but for production workloads geo-redundancy is recommended.

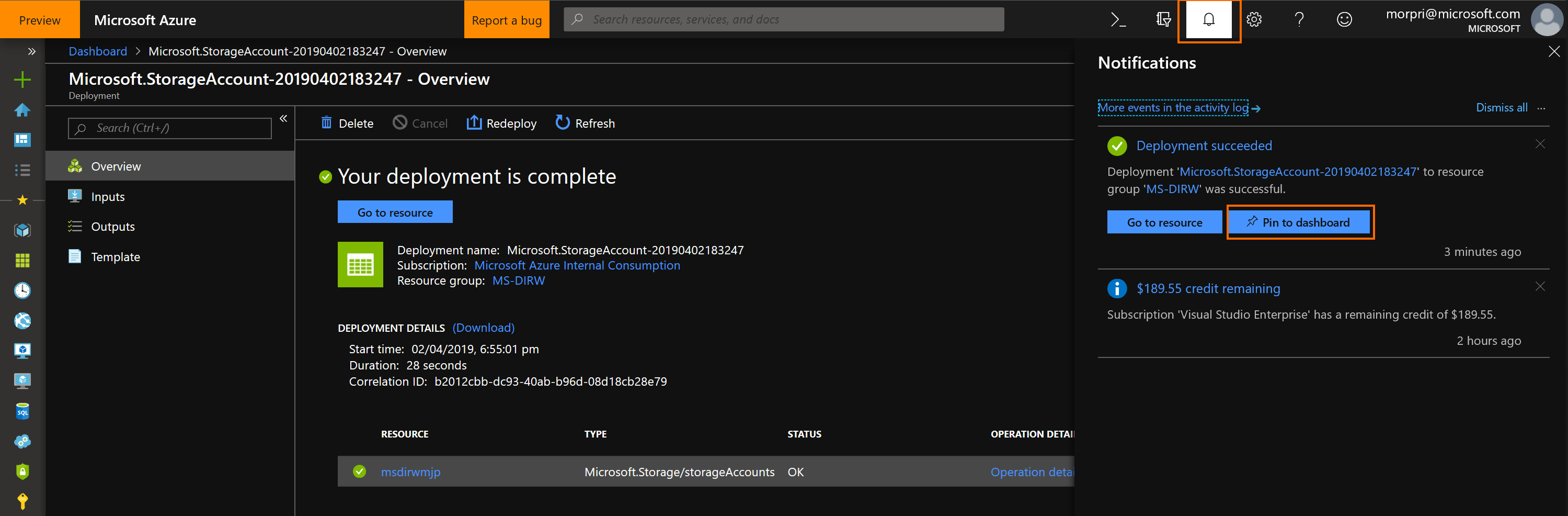
* 1. Access tier: “Hot”



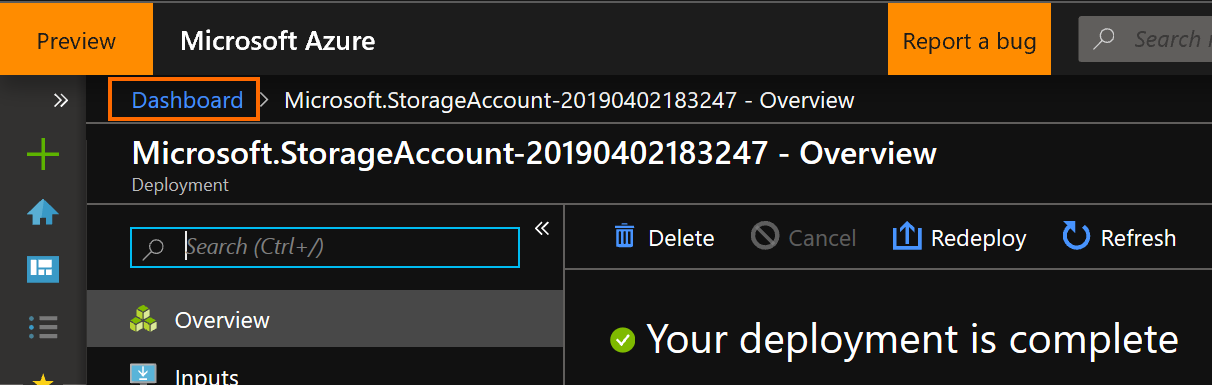
1. Under the “Advanced” tab we will use the defaults, note that we have enabled “Secure transfer” and that we can restrict access to certain networks to reduce surface area



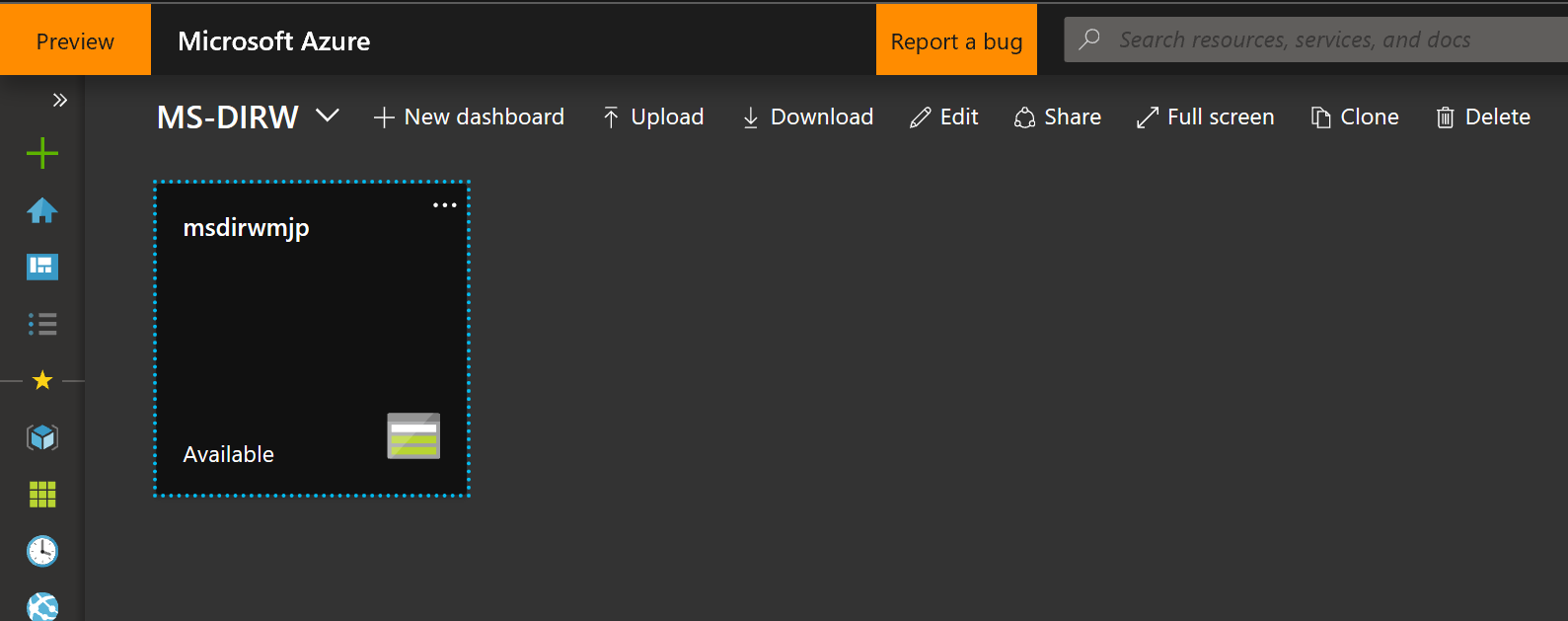
1. Review and Create your Storage account
2. When complete click on the notification (access using the little bell icon at the top right) and then click the “Pin to dashboard” button, this will make it easier to access this resource later



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”)

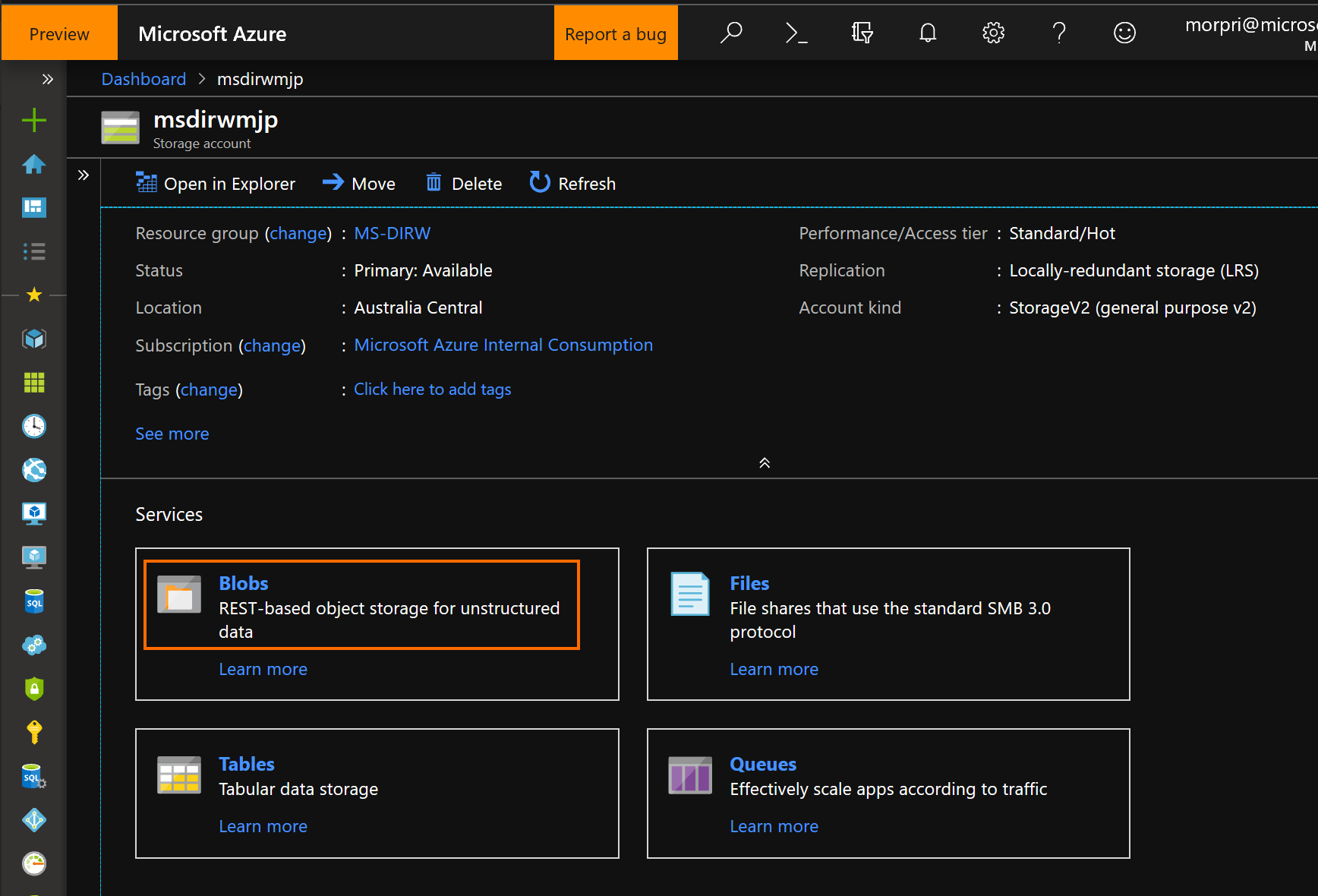


1. You should now see your new Storage account resource on your Dashboard

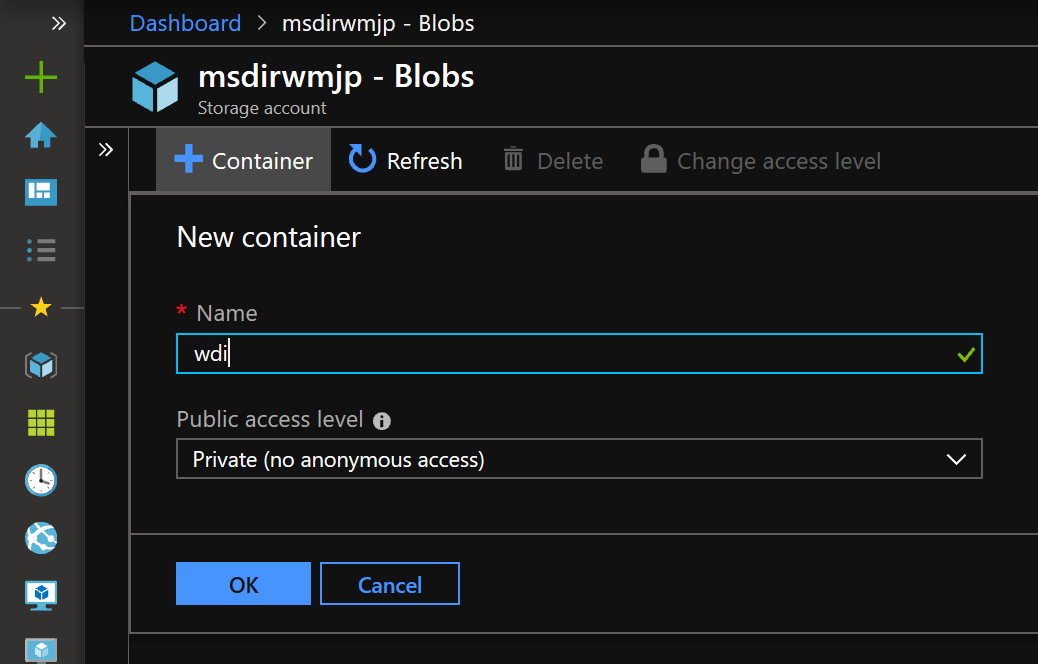


*Task 2: Create a new container to hold our raw data*

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



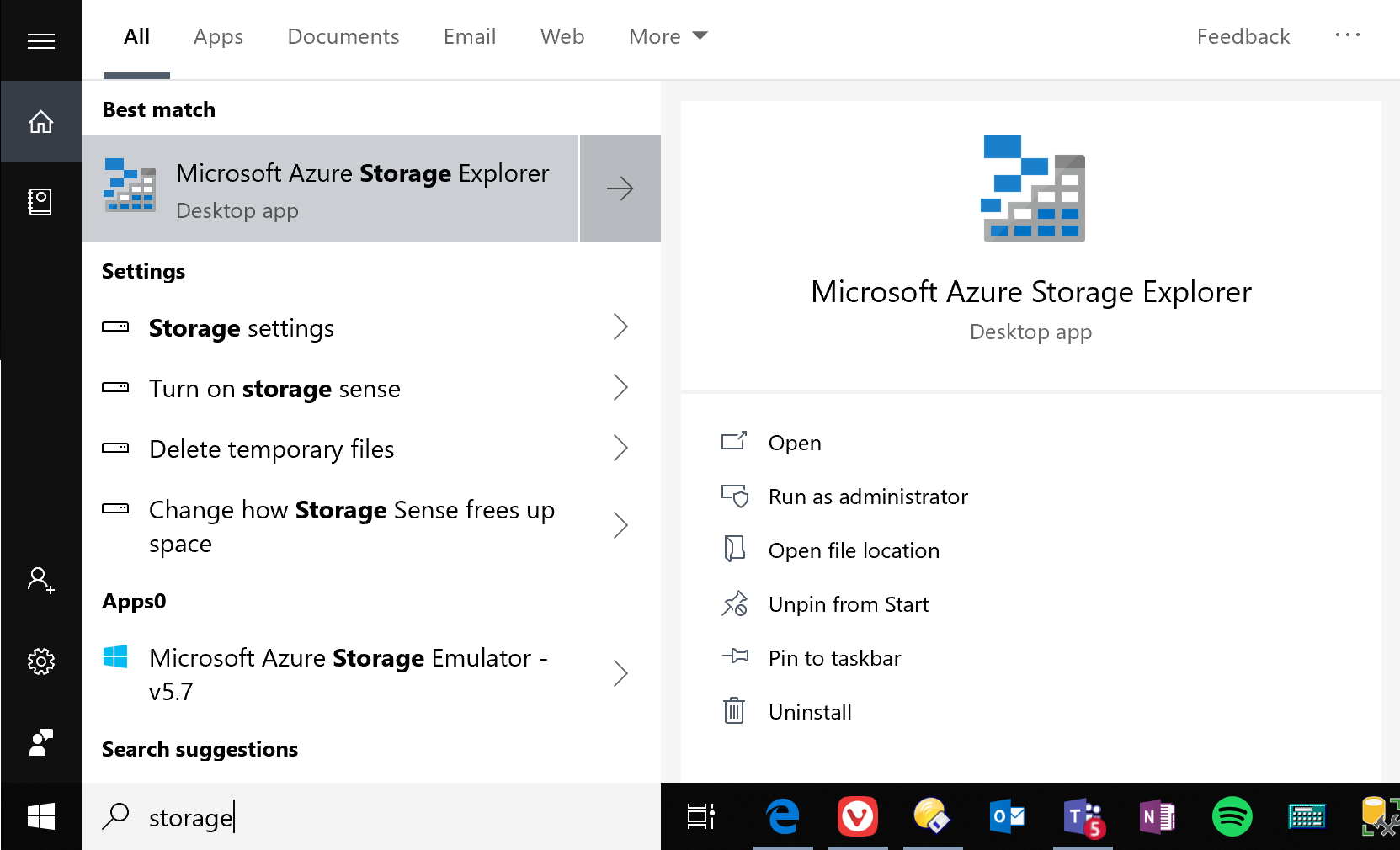
Exercise 3: Use Azure Storage Explorer (ASE) to Upload Raw Data

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

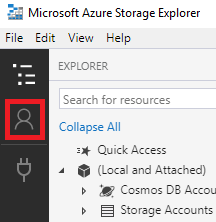
#### Task 1: Connect to an Azure subscription with ASE

The example used in this task is based on connecting a Visual Studio subscription, please substitute this with the subscription you are using for this workshop.

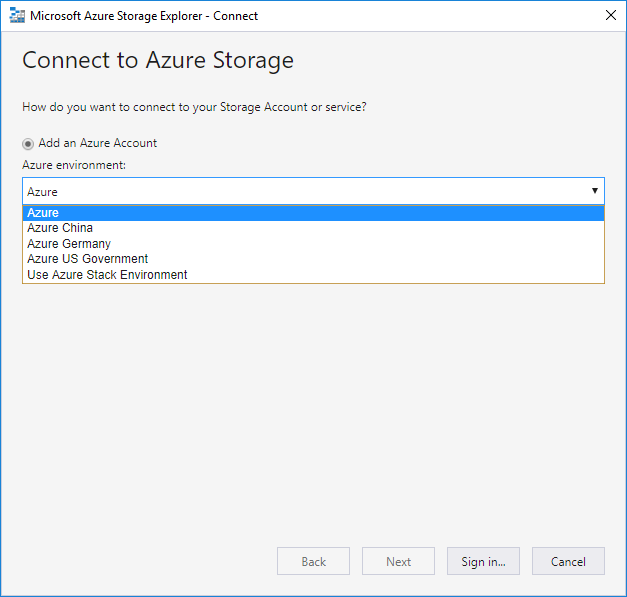
1. Start Storage Explorer



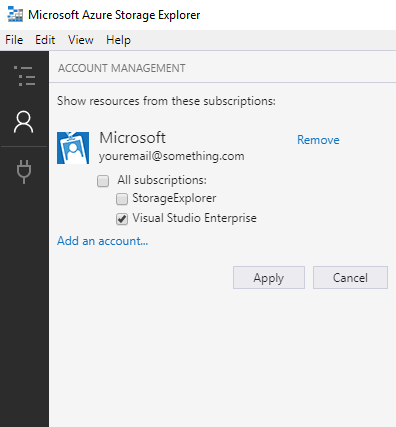
1. In Storage Explorer, select “Manage Accounts” to go to the “Account Management” panel.



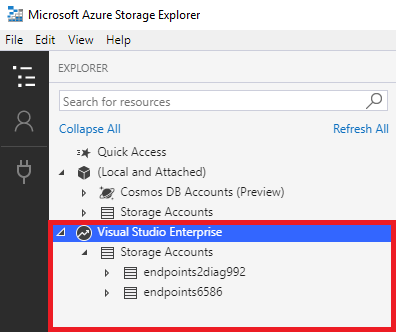
1. The left pane now displays all the Azure accounts you've signed in to. To connect to another account, select “Add an account”
2. Ensure “Azure” is selected in the environment dropdown



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)

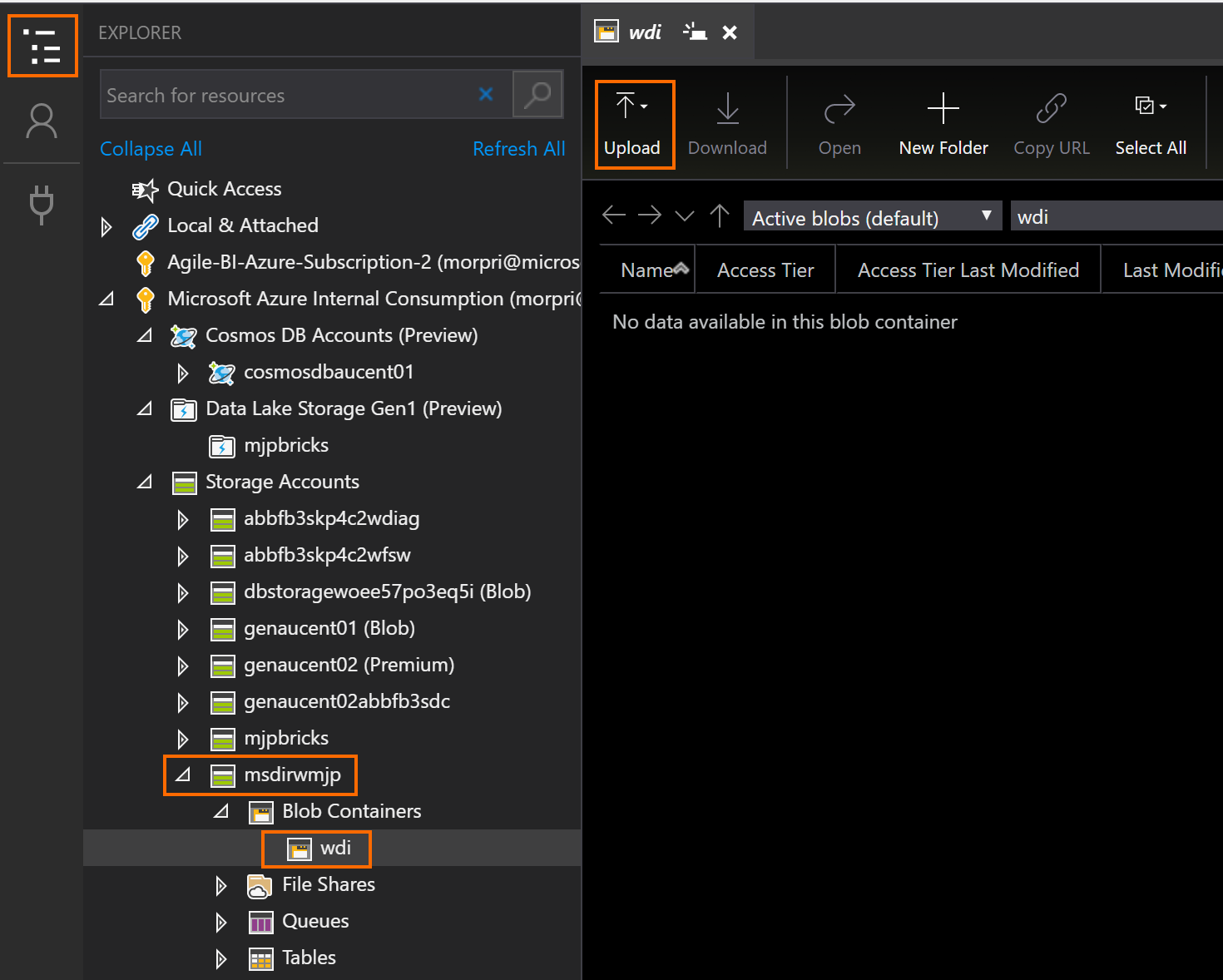


1. The left pane displays the storage accounts associated with the selected Azure subscriptions.

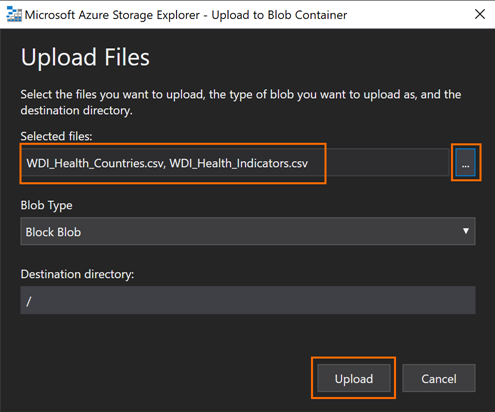


*Task 2: Upload the Raw CSV files*

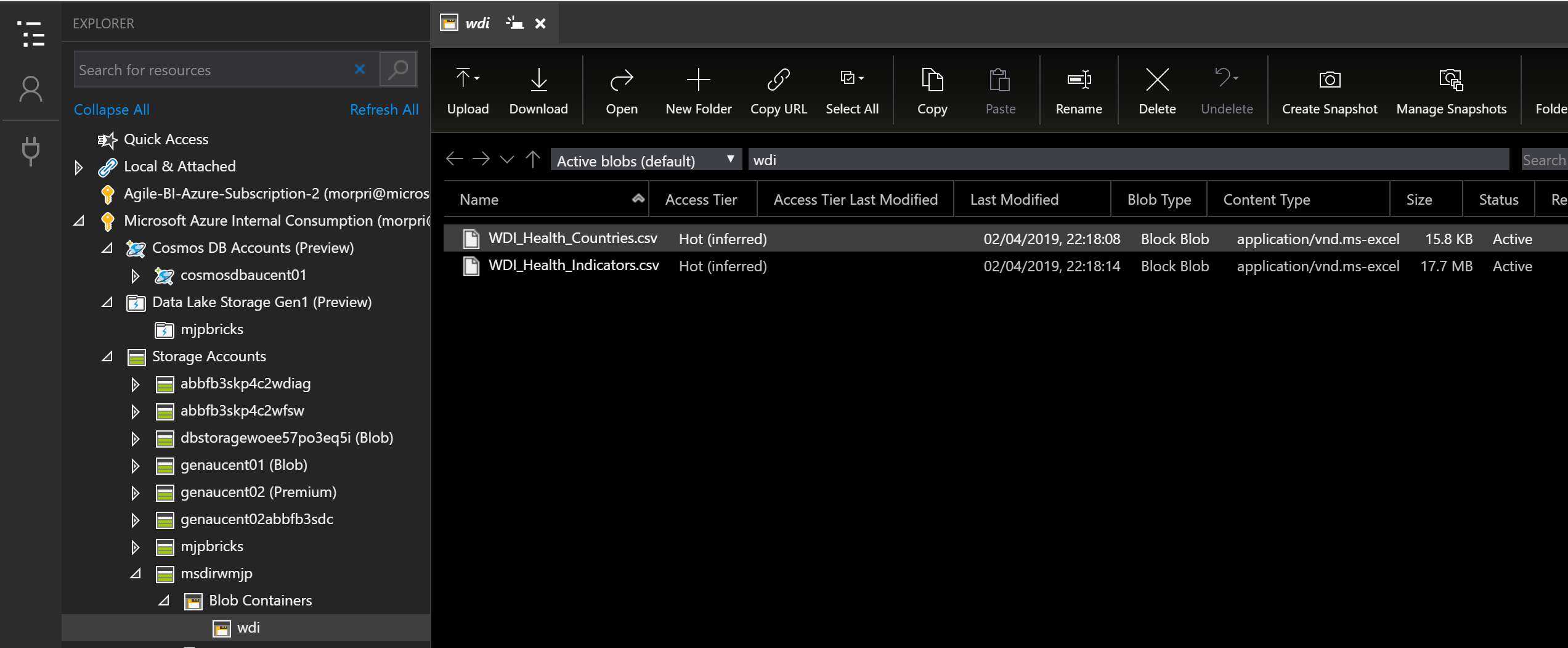
1. In Storage Explorer, navigate to your wdi container you created and select “Upload”



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:



## Lab 03: Deploy and Secure an Azure SQL Database Instance

In this lab we will deploy and configure an instance of an Azure SQL Database. We will then create the required artifacts (tables, views, users, etc.) to hold a structured version of our data to provide a data platform with the required performance and security for our analytics and reporting requirements.

Prerequisites

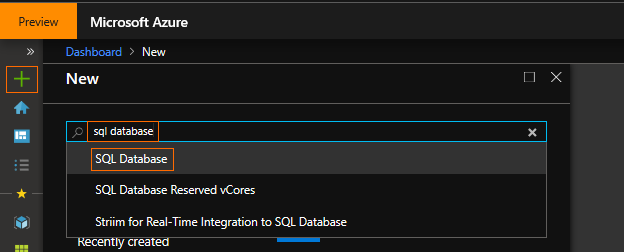
* Previous labs completed successfully
* Azure Data Studio installed

Exercise 1: Deploy an Azure SQL Server and Database Instance

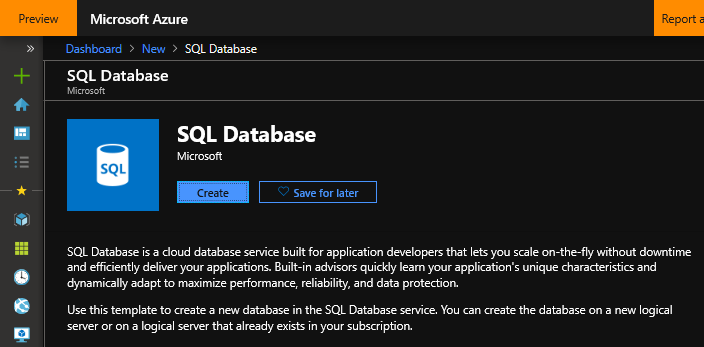
We will create a new server and deploy a database for our WDI data.

*Task 1: Deploy a new Azure SQL Database Instance*

1. Start from your MS-DIRW dashboard.
2. Click on “Create a resource”, type “sql database” in the search box and select the “SQL Database” entry.



1. Click “Create”



1. Fill in the following fields:
   1. Subscription: Select your subscription
   2. Resource group: Select the Resource Group you created when you created your Storage Account in Lab 02. The suggested name was “msdirw” appended with your initials, e.g. “msdirmjp”. Remember this name had to be globally unique.
   3. Database Name: “MS-DIRW”

Note that we are about to create a new logical server to host our database. A server can host multiple databases so we could use an existing database if we have one and can use this server in in the future to host new database instances.

* 1. Server: Click “Create new” and fill in the following fields:
     1. Server Name: Enter “msdirw” and append your initials, e.g. “msdirmjp”. This name must be globally unique so add additional information if your chosen name has already been used

Remember the name you have chosen so you can use this name to connect to the server later in the workshop.

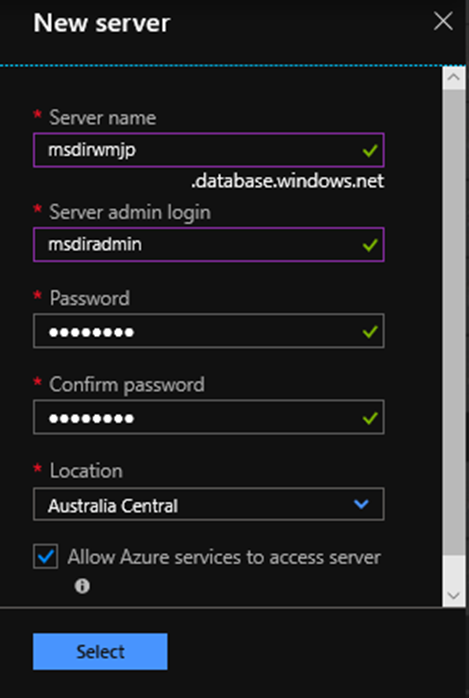
* + 1. Server admin login: “msdirwadmin”
    2. Password: “pa$$w0rd”

Of course, outside of this workshop please apply a far more robust and secure password policy. If you wish to use a different password now, please do so.

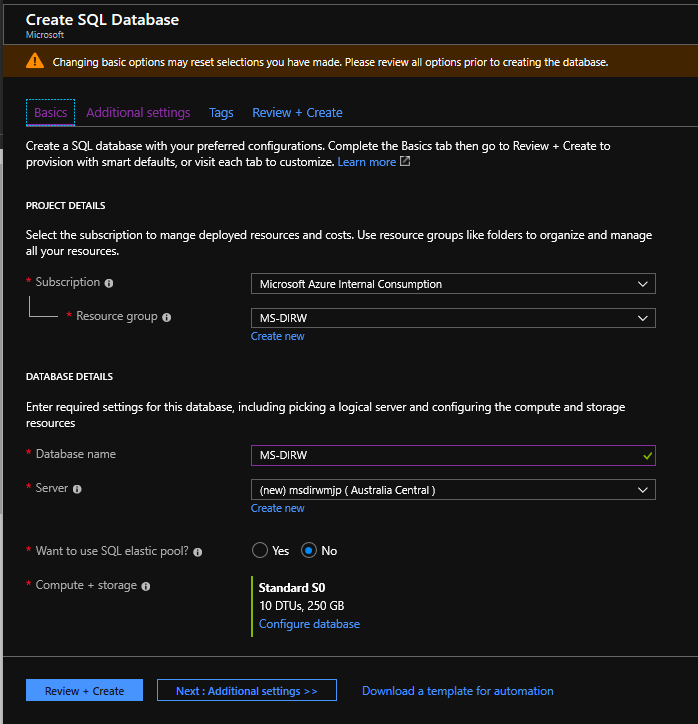
Remember to substitute “pa$$w0rd” with the one you have created whenever you see this password used in this workshop.

* + 1. Location: “Australia Central”

If you are using a different location for other resources deployed for this workshop use the same location for all resources for performance and cost reasons.



* + 1. Click “Select”
  1. Elastic pools: No
  2. Compute + storage: Standard S0, 10 DTUs, 250 GB



1. On the “Additional settings” tab fill in the following fields:
   1. Use existing data: None
   2. Collation: “Latin1\_General\_CI\_AS”
   3. Enable Advanced Data Security: “Start free trial”

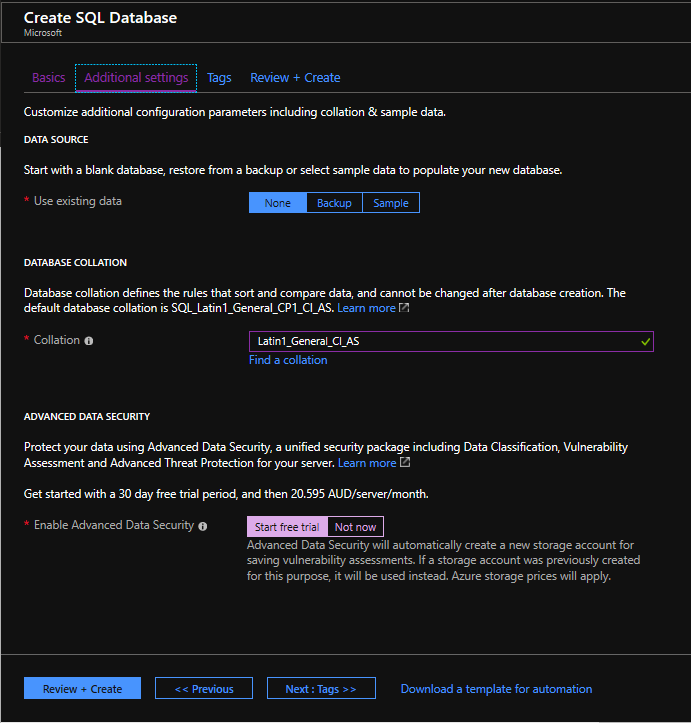
OPTIONAL only, enable this option to explore the advanced security features available with Azure SQL database such as:

**Advanced Treat Detection** – Get notified of security treats such as suspicious activity & connections, potentially harmful applications, SQL injection and brute force attacks.

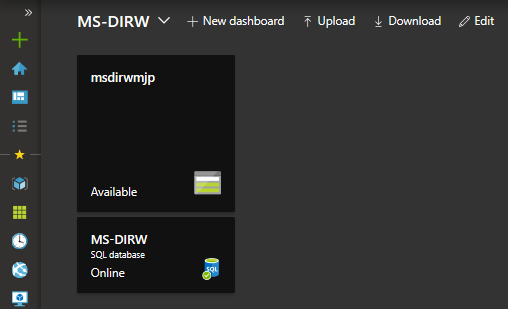
**Vulnerability Assessment** - Schedule regular assessments that check for insecure configuration, firewall rules, data classification recommendations, etc.

**Data Classification** – Classify data in your database so appropriate policy can be applied.

1. Click “Review + Create” and then click “Create” to deploy your new Azure SQL Database



1. 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 later
2. Click on the “Dashboard” link in the top left navigation path to get back to the dashboard
3. You should now see your new SQL Database on your Dashboard

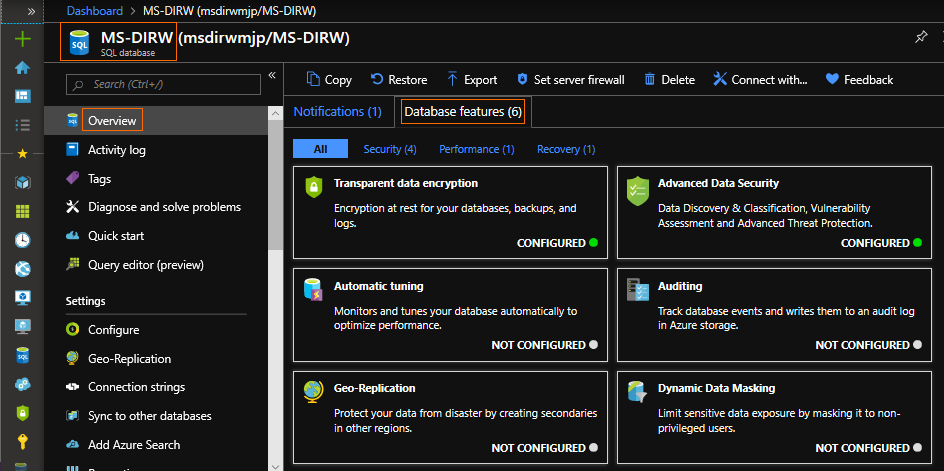


Exercise 3: Configure and Secure the SQL Database

We will ensure our database is secure and automate maintenance tasks.

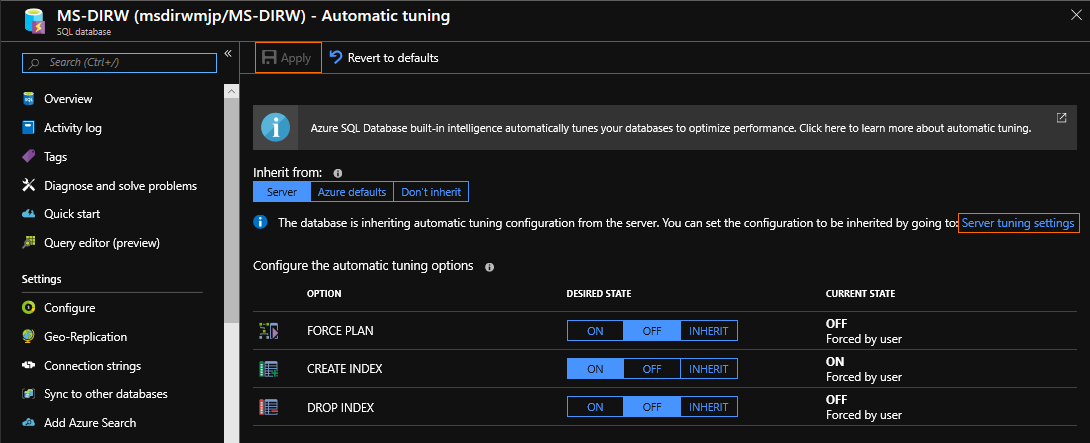
*Task 1: Configure Automatic tuning*

1. From the MS-DIRW dashboard select your SQL database.
2. At the bottom of the database Overview page select “Database features”, then select “Automatic tuning”



1. Configure the options as follows:
   1. FORCE PLAN: OFF, this is used to override join behavior of the query optimizer
   2. CREATE INDEX: ON, this will allow the database to create indexes to speed up data retrieval as it learns from how we access the data
   3. DROP INDEX: OFF, this will allow the database to drop any indexes that are not being used over time to free up space and improve insert and update performance

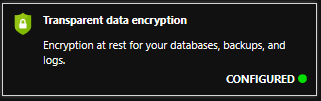
Alternatively, you can click on the “Server tuning settings” link and set these on the server, then set these to “INHERIT” for this database. This will then ensure all new databases created on the server will automatically get these settings as INHERIT is the default setting for new databases.



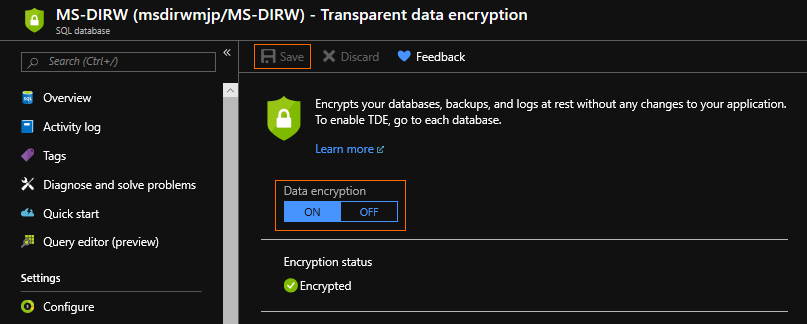
1. Click Apply

*Task 2: Configure Transparent Data Encryption (TDE)*

1. From the SQL database Overview page.
2. At the bottom of the database Overview page “Database features”, check if this database has already had “Transparent data encryption” configured by default, if not select “Transparent data encryption”, otherwise go to the next task. TDE will ensure that all data at rest, including backups and logs are encrypted.

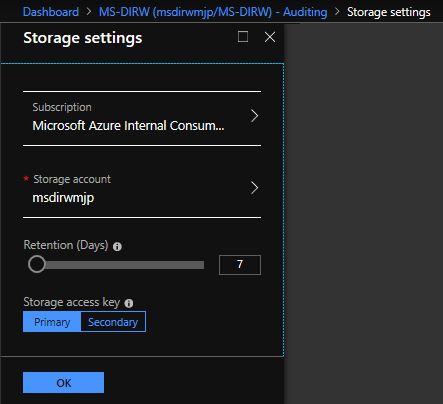


1. Select “ON” and then “Save”

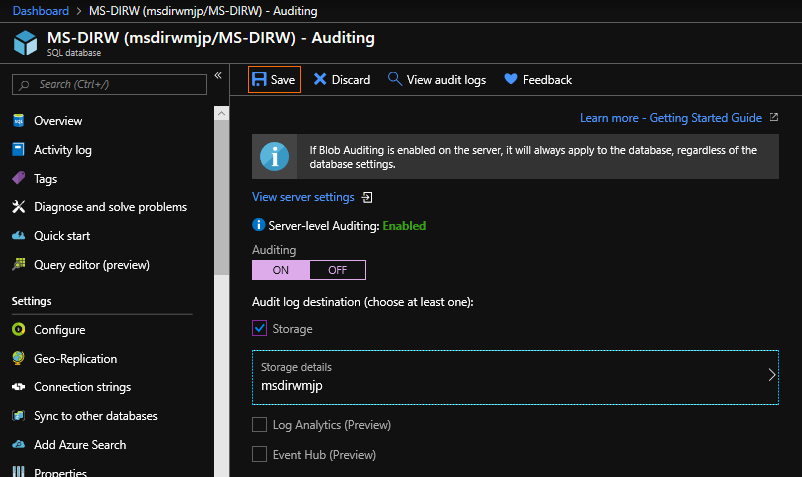


*Task 3: Configure database Auditing*

1. From the SQL database Overview page.
2. At the bottom of the database Overview page select “Database features”, then select “Auditing”
3. Fill in the following fields:
   1. Auditing: ON
   2. Audit log destination: Storage
   3. Storage Details as follows:
      1. Subscription: Select the subscription you are using for the workshop
      2. Storage account: Select the “msdirw” storage account you created (with the unique name)
      3. Retention: 7 days



* + 1. Click OK
  1. Click Save

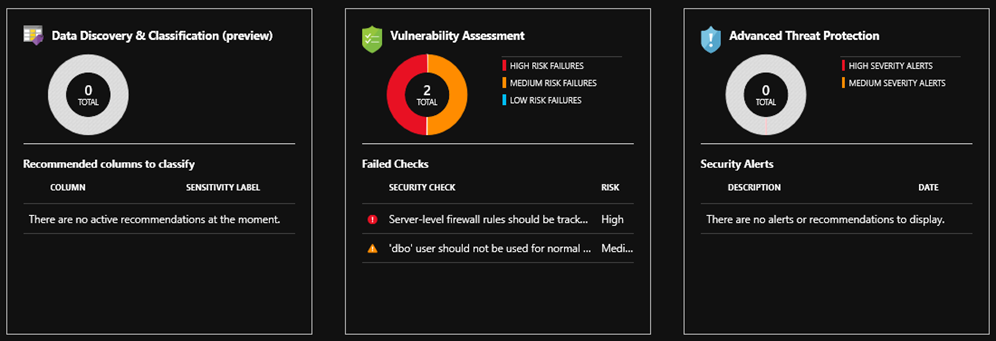


The remaining database features we will not be configuring are:

**Geo-Replication** - we will not need to configure this for our workshop however it is strongly recommended that this option be configured in production deployments to safeguard your database from events such as a regional disaster.

**Dynamic Data Masking** - this allows us to set rules to mask data from non-admin users, such as masking identification, credit cards, tax numbers, email addresses, etc.

**Advanced Data Security** - Advanced Treat Detection, Vulnerability Assessment and Data Classification

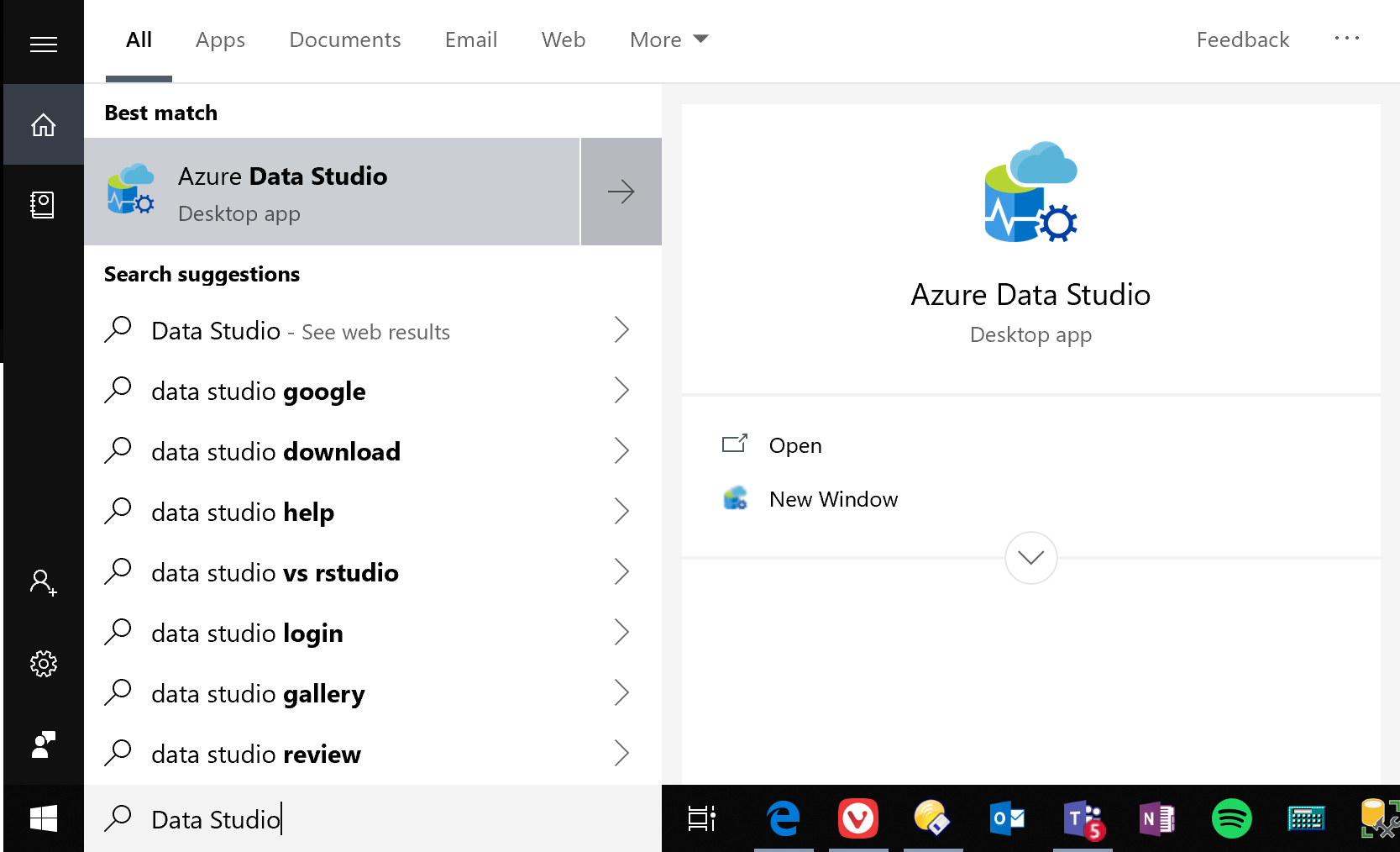


Exercise 3: Create the Azure SQL Database Artifacts to Hold WDI Data

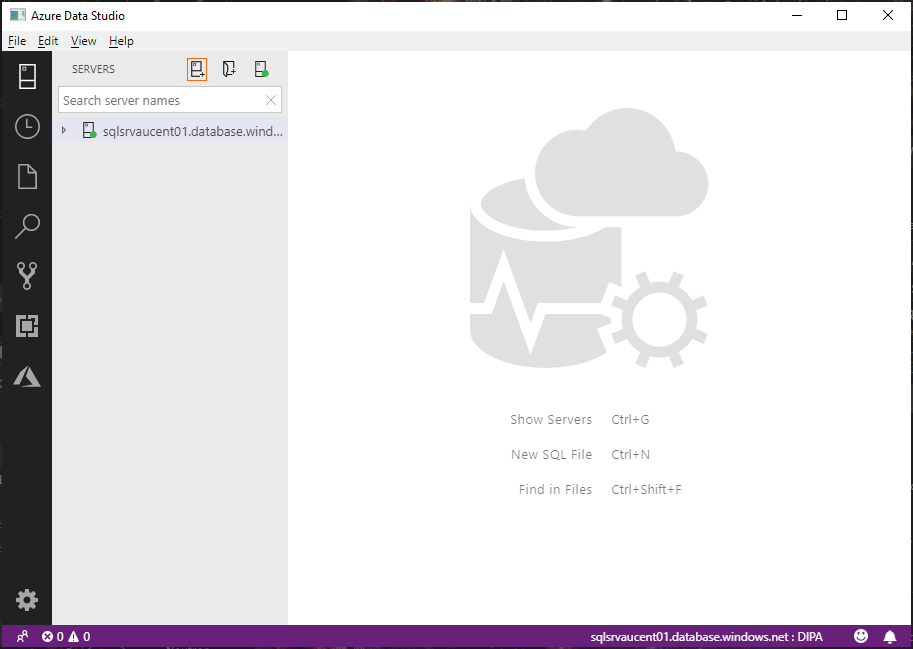
We will use a TSQL script to execute on our new database to create the tables and views we require for our data.

*Task 1: Execute a TSQL script against our database using* *Azure Data Studio*

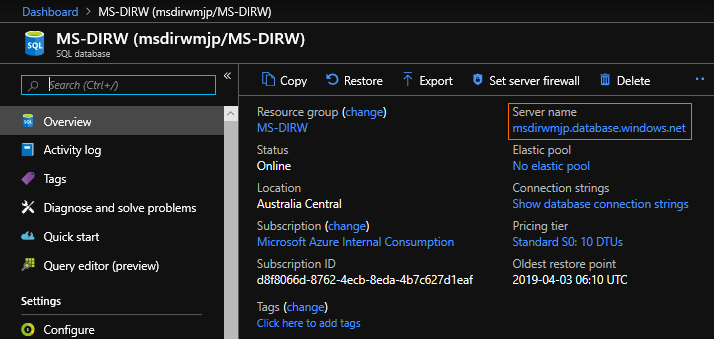
1. Start Azure Data Studio



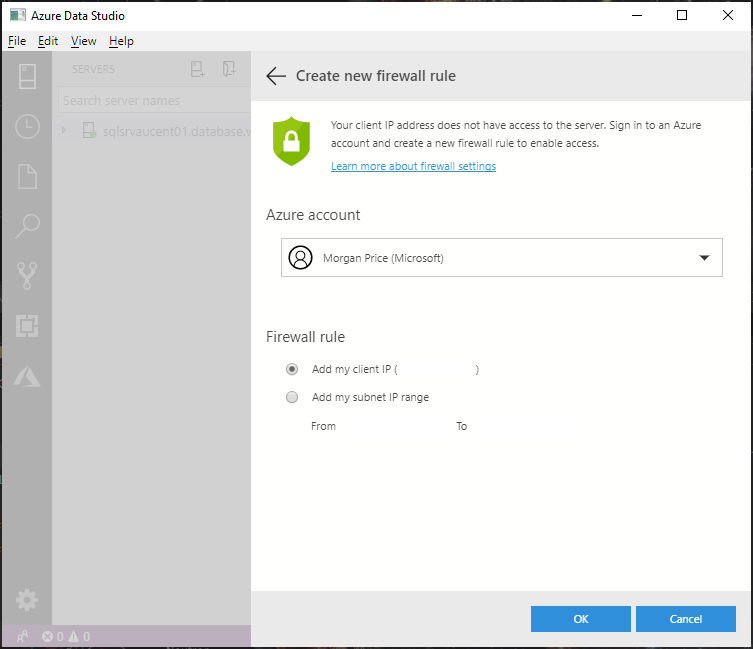
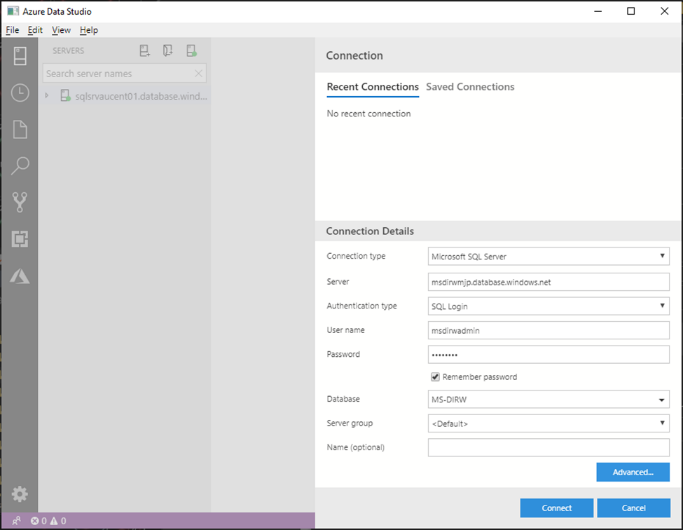
1. Click on the New Connection icon



1. Use the following connection details (we are using the Server Administrator credentials we created when we configured the server during the database deployment task):
   1. Connection type: Microsoft SQL Server
   2. Server: Copy your “Server name” from the SQL database Overview page



* 1. Authentication type: SQL Login
  2. User name: msdirwadmin
  3. Password: pa$$w0rd
  4. Remember password: tick
  5. Database: MS-DIRW, when you select this dropdown Azure Data Studio will connect to the database. You will be required to authenticate your account and will then be prompted to add your IP address to the server firewall.



1. Click “Connect”