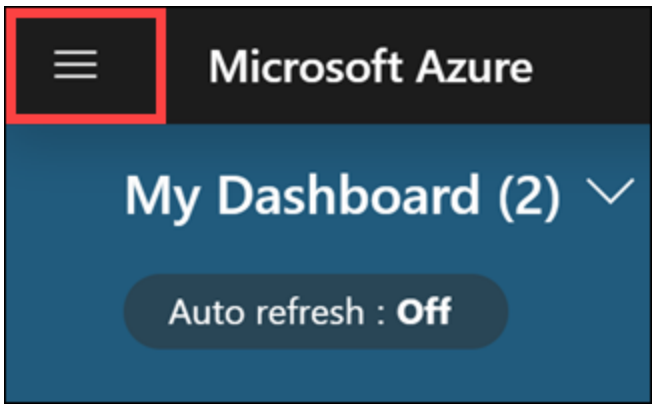
Before the hands-on lab

In this exercise, you will set up your environment for use in the rest of the hands-on lab. You should follow all the steps provided in the Before the Hands-on Lab section to prepare your environment *before* attending the hands-on lab.

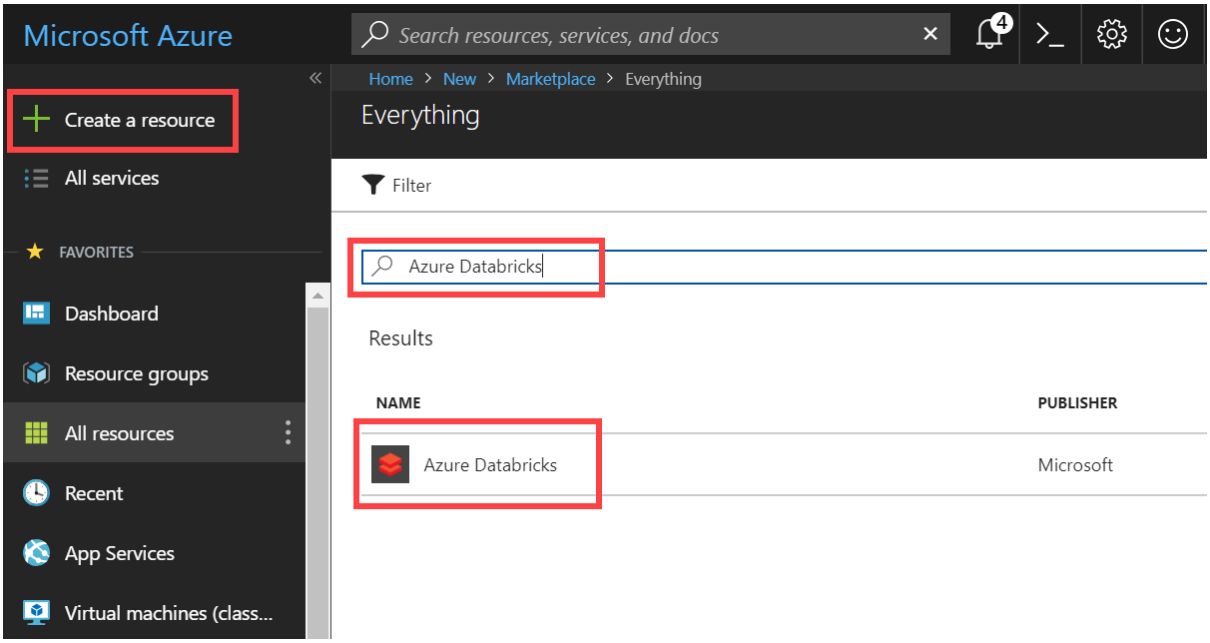
Task 1: Provision Azure Databricks

Azure Databricks is an Apache Spark-based analytics platform optimized for Azure. It will be used in this lab to build and train a machine learning model used to predict flight delays.

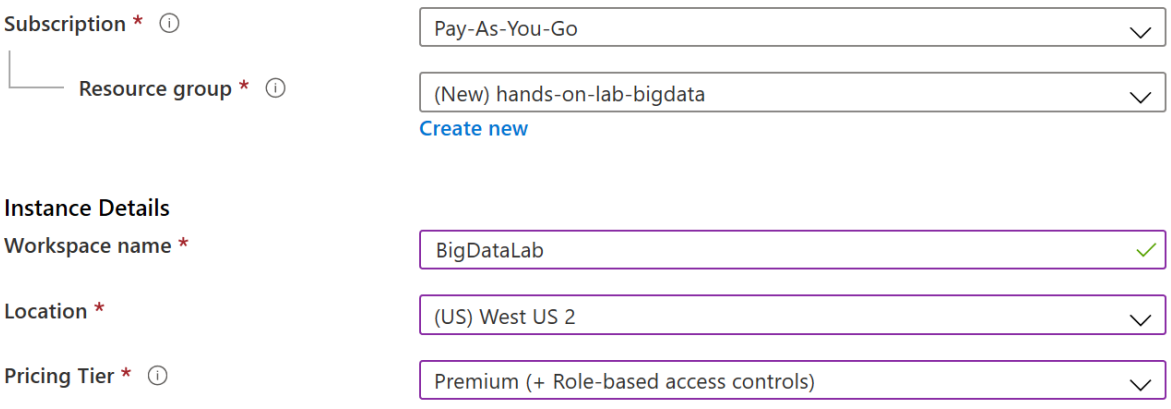
Note: To view the Azure portal menu, select the menu icon in the upper left-hand corner.



1. In the Azure Portal (https://portal.azure.com), select **+ Create a resource** within the portal menu, then type "Azure Databricks" into the search bar. Select Azure Databricks from the results.



2. Select **Create**.
3. Set the following configuration on the Azure Databricks Service creation form:
    - **Subscription**: Select the subscription you are using for this hands-on lab.
    - **Resource Group**: Select **Create new** and enter a unique name, such as `hands-on-lab-bigdata`
    - **Workspace name**: Enter a unique name, this is indicated by a green checkmark.
    - **Location**: Select a region close to you. ***(If you are using an Azure Pass, select South Central US.)***
    - **Pricing**: Select **Premium (+ Role-based access controls)**



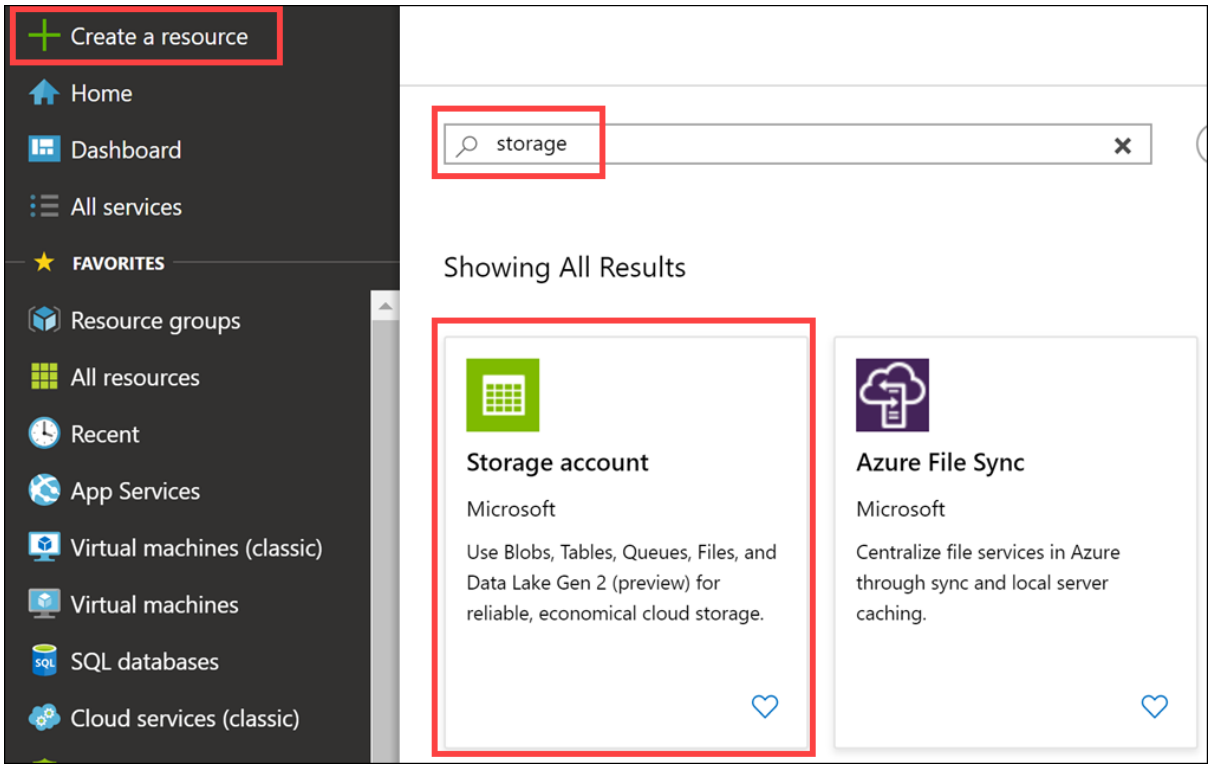
4. Select **Review + Create**.

5. Wait for validation to pass, then select **Create**.

Task 2: Create Azure Storage account

Create a new Azure Storage account that will be used to store historic and scored flight and weather data sets for the lab.

1. In the Azure Portal (https://portal.azure.com), select **+ Create a resource**, then type "storage" into the search bar. Select **Storage account** from the results.



2. Select **Create**.

3. Set the following configuration on the Azure Storage account creation form:

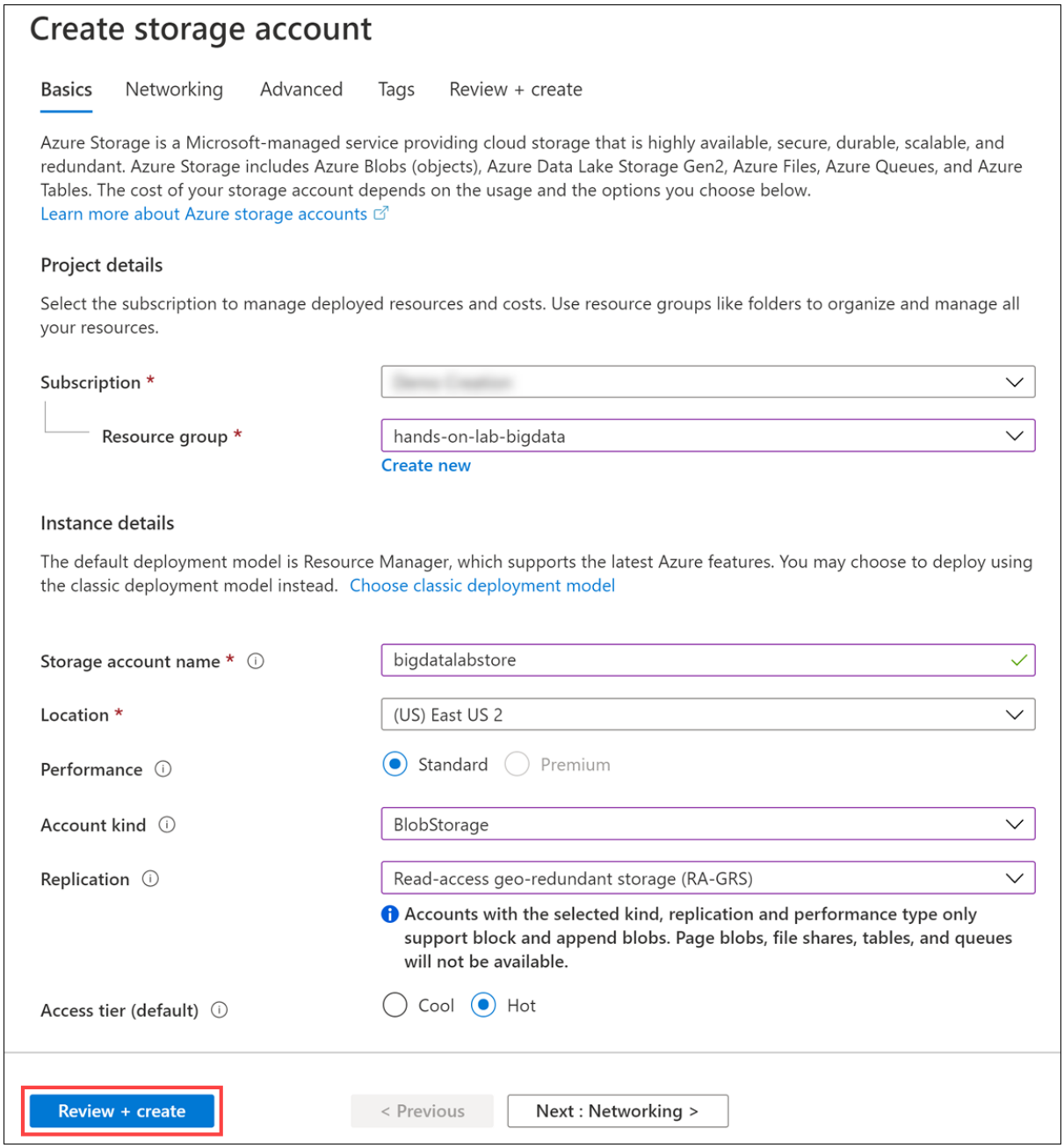
    - **Subscription**: Select the subscription you are using for this hands-on lab.

    - **Resource group**: Select the same resource group you created at the beginning of this lab.

    - **Storage account name**: Enter a unique name, this is indicated by a green checkmark.

    - **Location**: Select the same region you used for Azure Databricks.

- **Performance**: **Standard**
- **Account kind**: **BlobStorage**
- **Replication**: **Read-access geo-redundant storage (RA-GRS)**
- **Access tier**: **Hot**



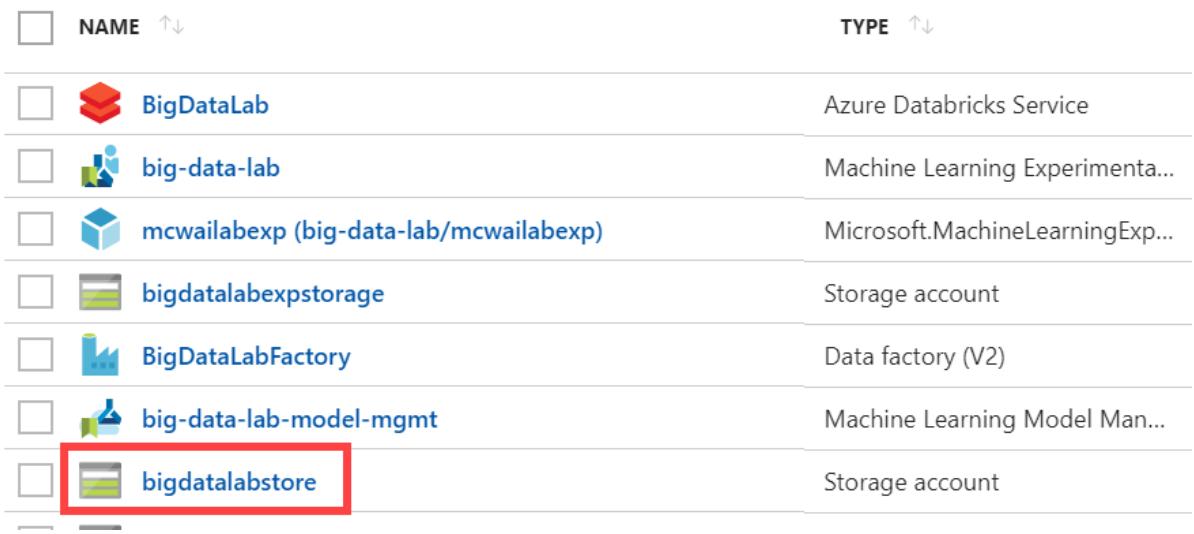
4. Select **Review + create**.
5. Wait for validation to pass, then select **Create**.

Task 3: Create storage container

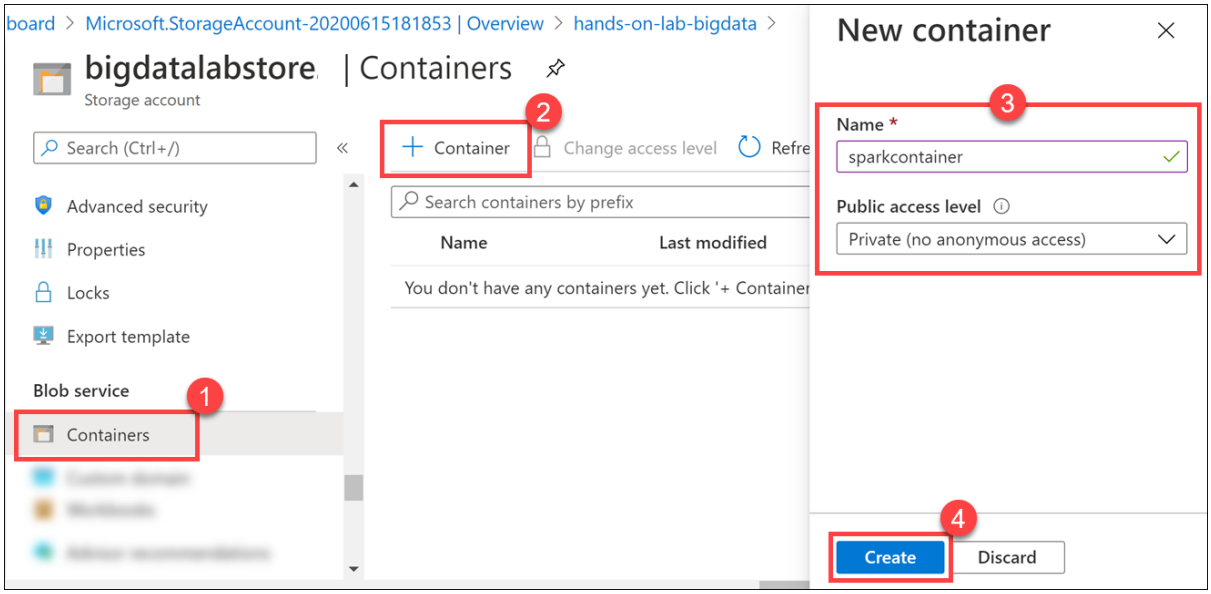
In this task, you will create a storage container in which you will store your flight and weather data files.

1. From the side menu in the Azure portal, choose **Resource groups**, then enter your resource group name into the filter box, and select it from the list.

2. Next, select your lab Azure Storage account from the list.



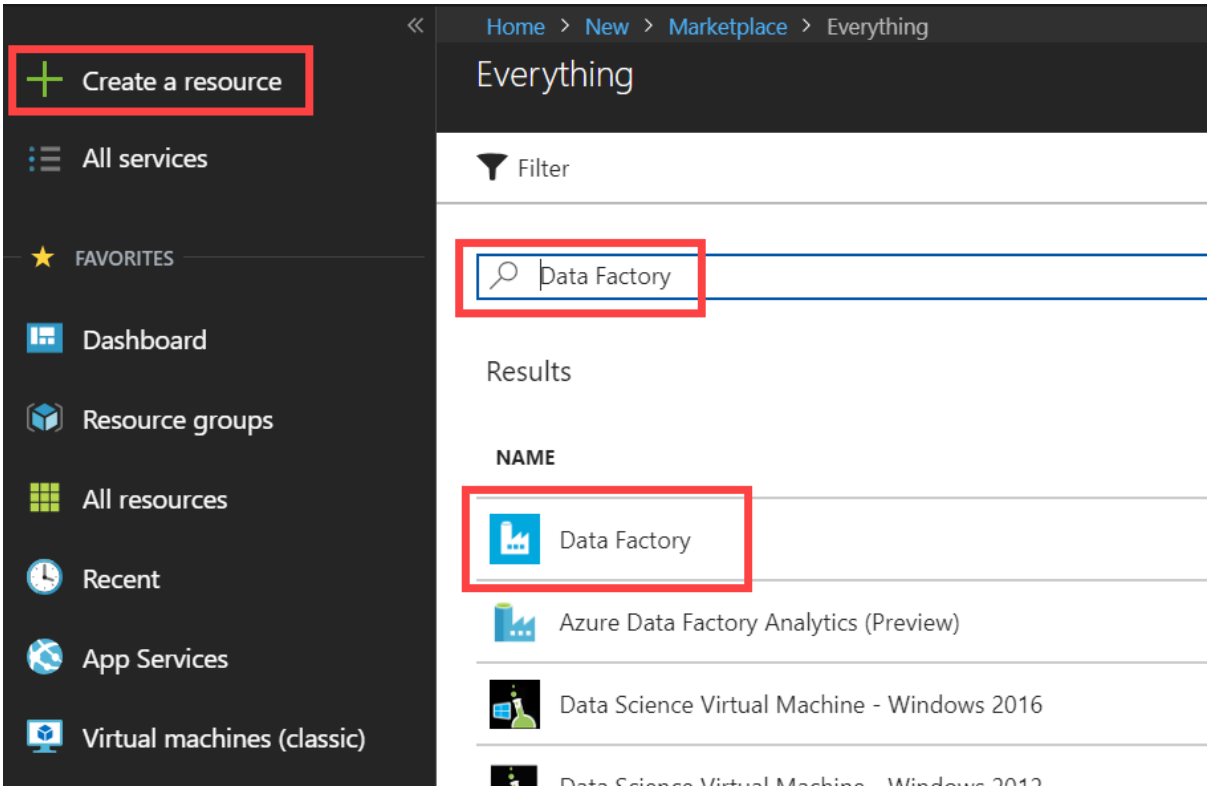
3. Select **Containers** (1) from the menu. Select **+ Container** (2) on the Containers blade, enter **sparkcontainer** for the name (3), leaving the public access level set to Private. Select **Create** (4) to create the container.



Task 4: Provision Azure Data Factory

Create a new Azure Data Factory instance that will be used to orchestrate data transfers for analysis.

1. In the Azure Portal (https://portal.azure.com), select **+ Create a resource**, then type "Data Factory" into the search bar. Select **Data Factory** from the results.



2. Select **Create**.

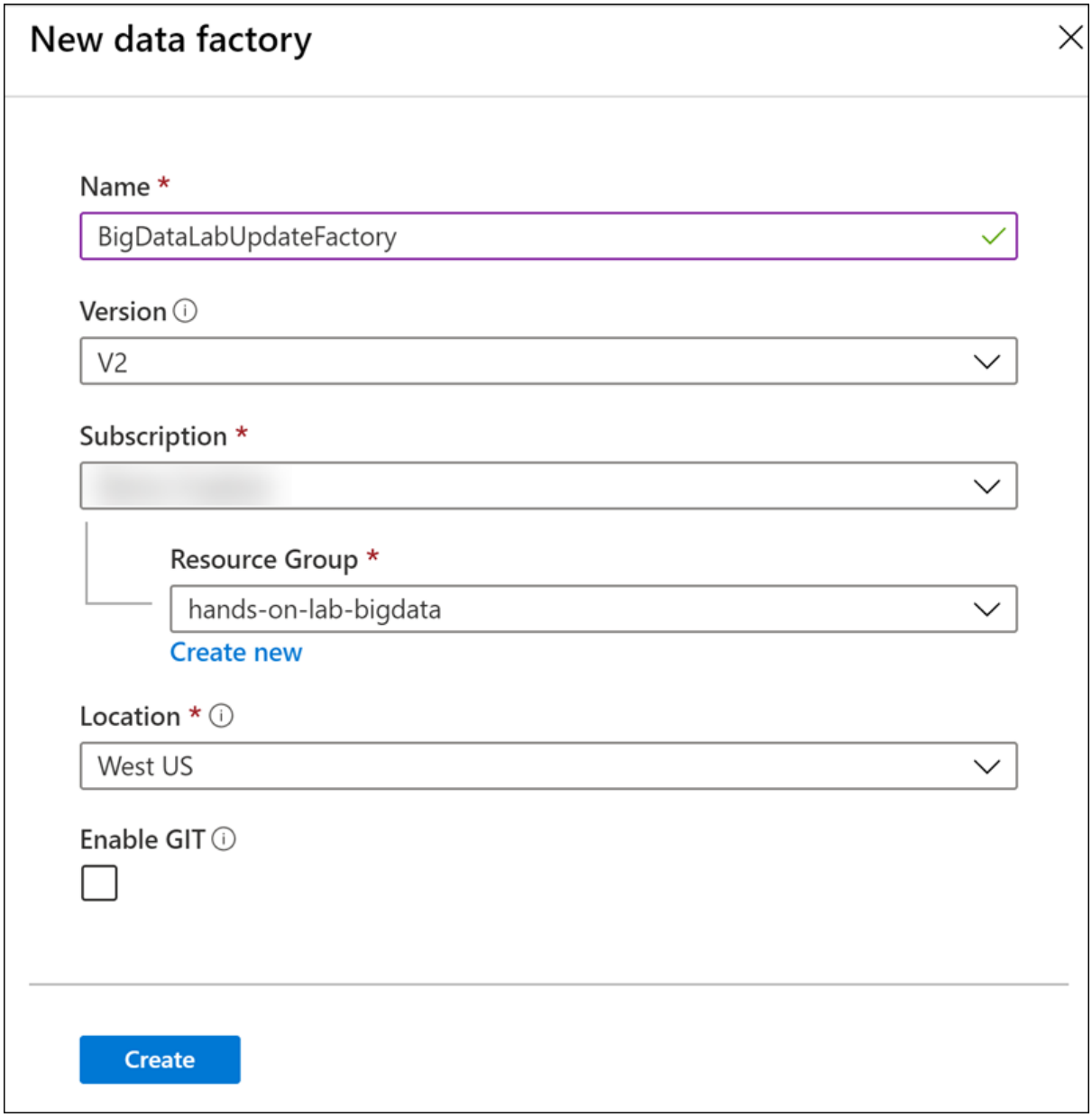
3. Set the following configuration on the Data Factory creation form:

    - **Name**: Enter a unique name, this is indicated by a green checkmark.
    - **Subscription**: Select the subscription you are using for this hands-on lab.
    - **Resource Group**: Select the same resource group you created at the beginning of this lab.
    - **Version**: Select **V2**
    - **Location**: Select any region close to you.
    - **Enable GIT**: **Unchecked**

    Understanding Data Factory Location: The Data Factory location is where the metadata of the data factory is stored and where the triggering of the pipeline is initiated from. Meanwhile, a data factory can access data stores and compute services in other Azure regions to move data between data stores or process data using compute services. This behavior is realized through

the globally available IR to ensure data compliance, efficiency, and reduced network egress costs.

The IR Location defines the location of its back-end compute, and essentially the location where the data movement, activity dispatching, and SSIS package execution are performed. The IR location can be different from the location of the data factory it belongs to.



4. Select **Create** to finish and submit.

Task 5: Download and install Power BI Desktop

Power BI desktop is required to make a connection to your Azure Databricks environment when creating the Power BI dashboard.

1. Download and install Power BI Desktop.

You should follow all these steps provided *before* attending the Hands-on lab.