

# Big data and visualization

---

Pre Deployment lab setup guide

## Contents

- [Big data and visualization before the hands-on lab setup guide](#)
  - [Requirements](#)
  - [Before the hands-on lab](#)
    - [Task 1: Provision Azure Databricks](#)
    - [Task 2: Create Azure Storage account](#)
    - [Task 3: Create a storage container](#)
    - [Task 4: Provision Azure Data Factory](#)
    - [Task 5: Download and install Power BI Desktop](#)
    - [Task 6: \(Optional\) Provision a VM to install the Integration Runtime On](#)

## Big data and visualization before the hands-on lab setup guide

---

### Requirements

---

1. Microsoft Azure subscription must be pay-as-you-go or MSDN.
  - Trial subscriptions will not work.
2. If you are not a Service Administrator or Co-administrator for the Azure subscription, or if you are running the lab in a hosted environment, you will need to install [Visual Studio 2019 Community](#) with the **ASP.NET and web development** and **Azure development** workloads.

### Before the hands-on lab

---

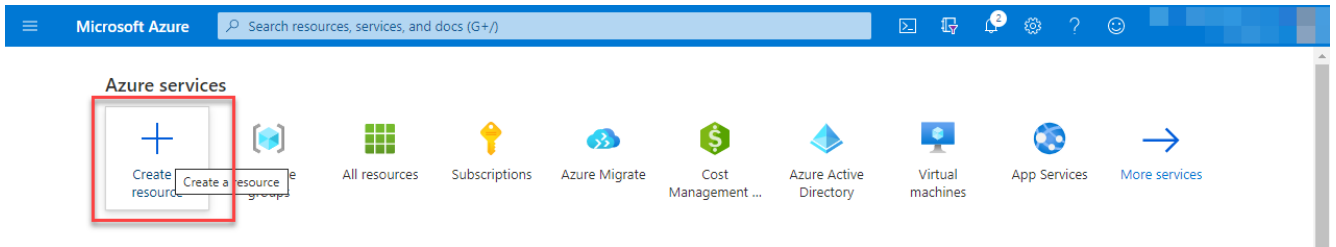
Duration: 30 minutes

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.

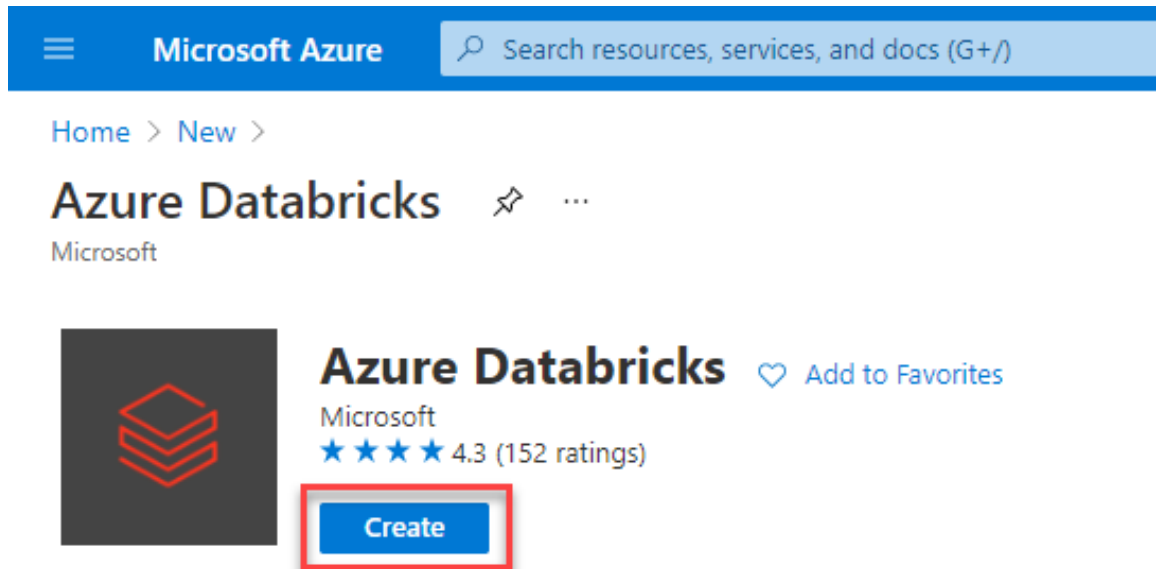
### Task 1: Provision Azure Databricks

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.

1. In the [Azure Portal](https://portal.azure.com) (<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 (1):** Select the subscription you are using for this hands-on lab.
- **Resource Group (2):** Select **Create new** and enter a unique name, such as `hands-on-lab-bigdata`.
- **Workspace name (3):** Enter a unique name, this is indicated by a green checkmark.
- **Location (4):** Select a region close to you. (*If you are using an Azure Pass, select South Central US.*)
- **Pricing (5):** Select **Premium (+ Role-based access controls)**

# Create an Azure Databricks workspace ...

Basics Networking Tags Review + create

## Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Demo Creation

Resource group \* ⓘ

hands-on-lab-bigdata

[Create new](#)

## Instance Details

Workspace name \*

BigDataLab

Region \*

West US 2

Pricing Tier \* ⓘ

Premium (+ Role-based access controls)

Review + create

< Previous

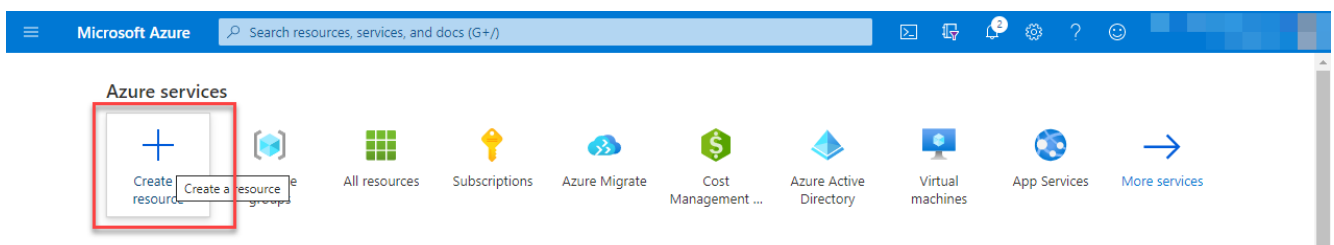
Next : Networking >

4. Select **Review + Create** (6).
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) (<https://portal.azure.com>), select **Create a resource** within the portal menu, then type "storage" into the search bar.



2. Select **Create** (2) and select **Storage Account** (3).

# Marketplace ...

The screenshot shows the Azure Marketplace interface. On the left is a sidebar with navigation links: Private Marketplace, Favorites, Recently created, Service Providers, Categories, Get Started (highlighted), AI + Machine Learning, Analytics, Blockchain, Compute, Containers, Databases, Developer Tools, DevOps, Identity, Integration, and Internet of Things. The main area has a search bar at the top with the text 'storage' and a red circle with the number '1' next to it. To the right of the search bar are filters: 'Pricing : All' and 'Operating System : All'. Below the search bar, it says 'Showing results for 'storage'' and 'Showing 1 to 20 of 996 results.' There are three product cards visible. The first card is for 'StoneFly Cloud Storage' by StoneFly, Inc., with a 'Free trial' badge. The second card is for 'Storage account' by Microsoft, with a 4.2 rating. The third card is for 'CloudLanes Cloud S Gateway' by CloudLanes. Each card has a 'Create' button. A red circle with the number '2' is next to the 'Create' button of the 'Storage account' card. A red circle with the number '3' is next to the 'Storage account' card. Below the 'Storage account' card, there is a dropdown menu with the text 'Storage account' and a green checkmark.

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

- **Subscription (1):** Select the subscription you are using for this hands-on lab.
- **Resource group (2):** Select the same resource group you created at the beginning of this lab.
- **Storage account name (3):** Enter a unique name, this is indicated by a green checkmark.
- **Location (4):** Select the same region you used for Azure Databricks.
- **Performance (5): Standard**
- **Account kind (6): StorageV2**
- **Replication (7): Read-access geo-redundant storage (RA-GRS)**

## Create storage account ...

[Basics](#) [Networking](#) [Data protection](#) [Advanced](#) [Tags](#) [Review + create](#)

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below.

[Learn more about Azure storage accounts](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \*

Demo Creation

Resource group \*

hands-on-lab-bigdata

[Create new](#)

### Instance details

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

Storage account name \* ⓘ

bigdatalabstore10

Location \*

(US) West US 2

Performance ⓘ

☒ Standard ☐ Premium

Account kind ⓘ

StorageV2 (general purpose v2)

Replication ⓘ

Read-access geo-redundant storage (RA-GRS)

[Review + create](#)

[< Previous](#)

[Next : Networking >](#)

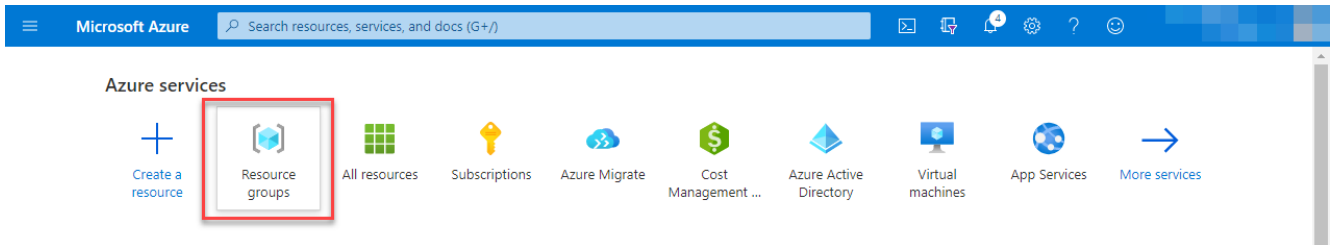
**Note:** There may be changes in the Azure UI. If you do not see the option to provision **RA-GRS** replication, choose **GRS** and select **Make read access to data available in the event of regional unavailability**. You may also not have the option to select the storage account type.

4. Select **Review + create (8)**.
5. Wait for validation to pass, then select **Create**.

## Task 3: Create a storage container

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

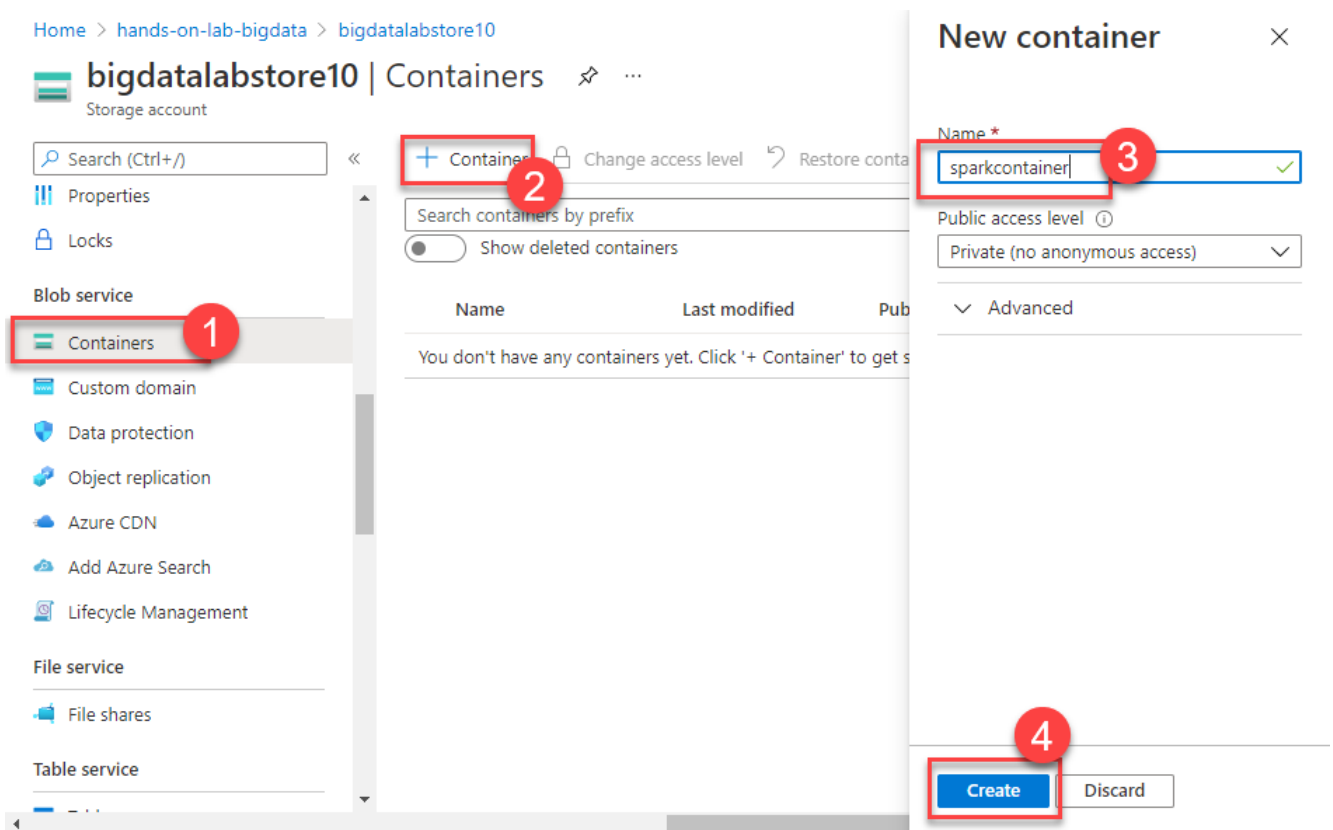
1. From the home page in the Azure portal, choose **Resource groups**, then enter your resource group name into the filter box, and select it from the list.



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

Name ↑↓	Type ↑↓	Location ↑↓
<input type="checkbox"/> BigDataLab	Azure Databricks Service	West US 2
<input type="checkbox"/> bigdatalabstore10	Storage account	West US 2

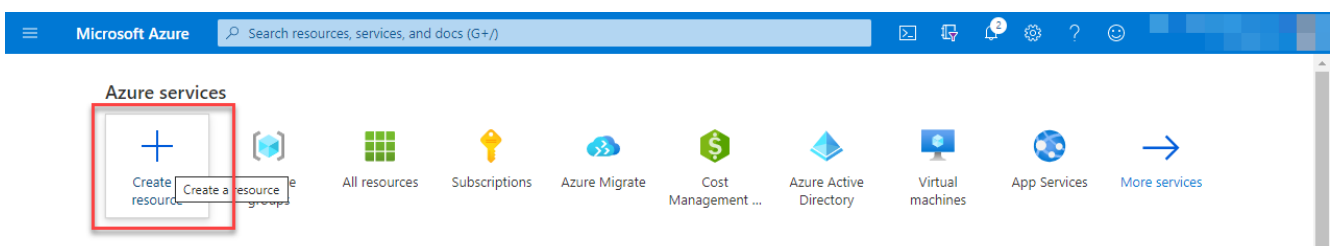
3. Select **Containers** (1) from the menu. Select **+ Container** (2) on the Containers blade, enter **sparkcontainer** for the name (3), leaving the public access level set to Private. Select **Create** (4) to create the container.



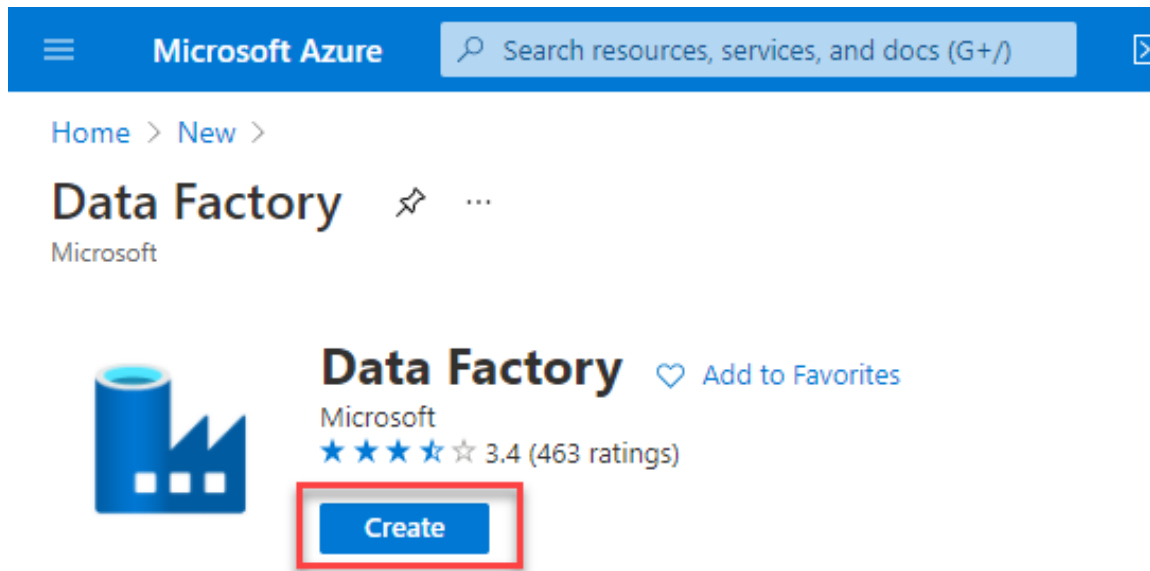
## Task 4: Provision Azure Data Factory

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) (<https://portal.azure.com>), select **Create a resource** within the portal menu, then type "Data Factory" into the search bar.



2. Select **Create**.



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

- **Subscription (1):** Select the subscription you are using for this hands-on lab.
- **Resource Group (2):** Select the same resource group you created at the beginning of this lab.
- **Region (3):** Select any region close to you.
- **Name (4):** Enter a unique name, this is indicated by a green checkmark.
- **Version (5):** Select **V2**

***Understanding Data Factory Location:***

The Data Factory location is where the data factory's metadata 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.

## Create Data Factory ...

Basics   Git configuration   Networking   Advanced   Tags   Review + create

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Demo Creation

1

Resource group \* ⓘ

hands-on-lab-bigdata

2

[Create new](#)

### Instance details

Region \* ⓘ

West US 2

3

Name \* ⓘ

bigdatalabdatafactory10

4

Version \* ⓘ

V2

5

**Review + create**

< Previous

Next : Git configuration >

6

4. Select **Next: Git configuration > (6)** to continue.
5. Check **Configure Git later (1)** and select **Review + create (2)** to proceed.

## Create Data Factory ...

×

Basics   Git configuration   Networking   Advanced   Tags   Review + create

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration.

[Learn more about Git integration in Azure Data Factory](#)

Configure Git later ⓘ



1

2

**Review + create**

< Previous

Next : Networking >

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



## Task 5: Download and install Power BI Desktop

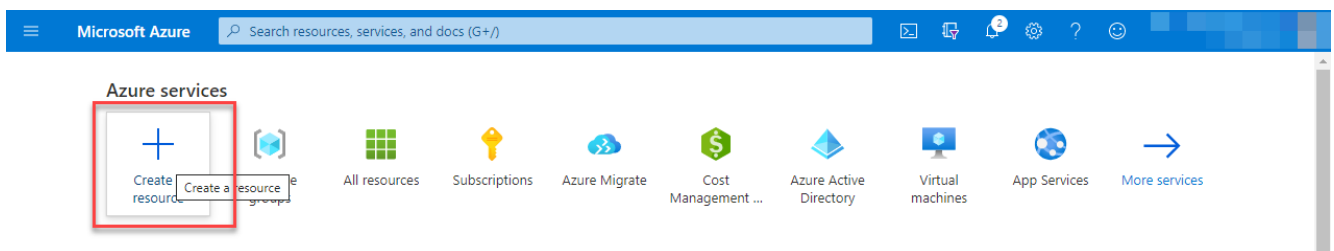
Power BI desktop is required to connect to your Azure Databricks environment when creating the Power BI dashboard.

1. Download and install [Power BI Desktop](#).

## Task 6: (Optional) Provision a VM to install the Integration Runtime On

An integration runtime agent for Azure Data Factory will need to be installed on your hardware for the hands-on lab. Since you will need to provide your user credentials, we suggest you provision an Azure VM to act as your "on-premises" hardware.

1. In the [Azure Portal](https://portal.azure.com) (<https://portal.azure.com>), select **Create a resource** within the portal menu.



2. Select **Windows Server 2016 Datacenter** from Azure Marketplace.

# New

 Search the Marketplace

Azure Marketplace [See all](#)

Popular

Get started

Recently created

AI + Machine Learning

Analytics

Blockchain

Compute

Containers

Databases

Developer Tools

DevOps

Identity

Integration

Internet of Things

IT & Management Tools

Media

Migration

Mixed Reality

Monitoring & Diagnostics

Networking

Security

Software as a Service (SaaS)



**Windows Server 2016 Datacenter**

[Quickstarts + tutorials](#)



**Ubuntu Server 18.04 LTS**

[Learn more](#)



**Web App**

[Quickstarts + tutorials](#)



**SQL Database**

[Quickstarts + tutorials](#)



**Function App**

[Quickstarts + tutorials](#)



**Azure Cosmos DB**

[Quickstarts + tutorials](#)



**Kubernetes Service**

[Quickstarts + tutorials](#)



**DevOps Starter**

[Quickstarts + tutorials](#)



**Storage account**

[Quickstarts + tutorials](#)

[Show recently created items](#)

3. On the **Create a virtual machine** page, specify the following parameters:
  - **Subscription (1)**: Provide the subscription you have been using for this lab.
  - **Resource group (2)**: Provide your resource group.
  - **Virtual machine name (3)**: Provide something descriptive.
  - **Region (4)**: Provide the same location as your ADF instance.

- **Availability options:** No infrastructure redundancy required
- **Image:** Windows Server 2016 Datacenter - Gen1
- **Azure Spot instance:** Unselected
- **Size (5):** Standard\_D2s\_v3
- **Username (6):** demouser
- **Password/Confirm password (6):** Password.1!!
- **Public inbound ports:** Allow selected ports
- **Select inbound ports:** RDP (3389)
- **Would you like to use an existing Windows Server license?** No

## Create a virtual machine ...

[Basics](#)   [Disks](#)   [Networking](#)   [Management](#)   [Advanced](#)   [Tags](#)   [Review + create](#)

Create a virtual machine that runs Linux or Windows. Select an image from Azure marketplace or use your own customized image. Complete the Basics tab then Review + create to provision a virtual machine with default parameters or review each tab for full customization. [Learn more](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Demo Creation

Resource group \* ⓘ

hands-on-lab-bigdata

[Create new](#)

### Instance details

Virtual machine name \* ⓘ

integrationruntime

Region \* ⓘ

(US) West US 2

Availability options ⓘ

No infrastructure redundancy required

Image \* ⓘ

Windows Server 2016 Datacenter - Gen1

[See all images](#)

Azure Spot instance ⓘ

☐

Size \* ⓘ

Standard\_D2s\_v3 - 2 vcpus, 8 GiB memory (\$137.24/month)

[See all sizes](#)

### Administrator account

Username \* ⓘ

demouser

Password \* ⓘ

.....

Confirm password \* ⓘ

.....

### Inbound port rules

Select which virtual machine network ports are accessible from the public internet. You can specify more limited or granular network access on the Networking tab.

Public inbound ports \* ⓘ

☐ None

☒ Allow selected ports

Select inbound ports \*

RDP (3389)

**⚠ This will allow all IP addresses to access your virtual machine.** This is only recommended for testing. Use the Advanced controls in the Networking tab to create rules to limit inbound traffic to known IP addresses.

**Review + create**

< Previous

Next : Disks >

4. select **Review + create (7)** to proceed.

5. Select **Create** on the validation page to finish and start provisioning your VM. When the deployment is complete **(1)**, select **Go to resource (2)** to navigate to your VM.

CreateVm-MicrosoftWindowsServer.WindowsServer-201-20210318194824 | Overview

Deployment

Search (Ctrl+/) << Delete Cancel Redeploy Refresh

We'd love your feedback! →

**✓ Your deployment is complete (1)**

Deployment name: CreateVm-MicrosoftWindowsServer.WindowsSe... Start time: 3/18/2021, 7:54:19 PM  
Subscription: Demo Creation Correlation ID: 3edfccfd-ac17-4941-a9b7-8e7cd71aca8f  
Resource group: hands-on-lab-bigdata

Deployment details (Download)

Next steps

Setup auto-shutdown Recommended  
Monitor VM health, performance and network dependencies Recommended  
Run a script inside the virtual machine Recommended

**2 Go to resource** Create another VM

Security Center  
Secure your apps and infrastru  
[Go to Azure security center >](#)

Free Microsoft tutorials  
[Start learning today >](#)

Work with an expert  
Azure experts are service provi  
who can help manage your as:  
and be your first line of suppo

6. Select **Connect** from the upper left-hand corner of the page. Then, select **RDP**. Finally, select **Download RDP File**.

**RDP** SSH BASTION

### Connect with RDP

To connect to your virtual machine via RDP, select an IP address, optionally change the port number, and download the RDP file.

IP address \*

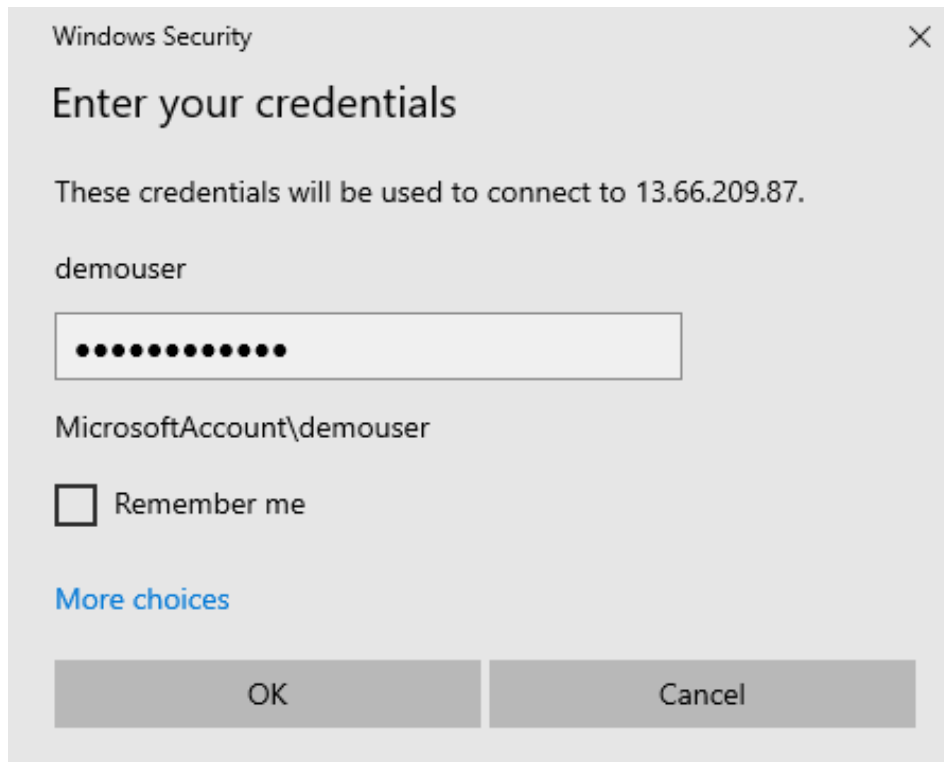
Public IP address (13.66.209.87)

Port number \*

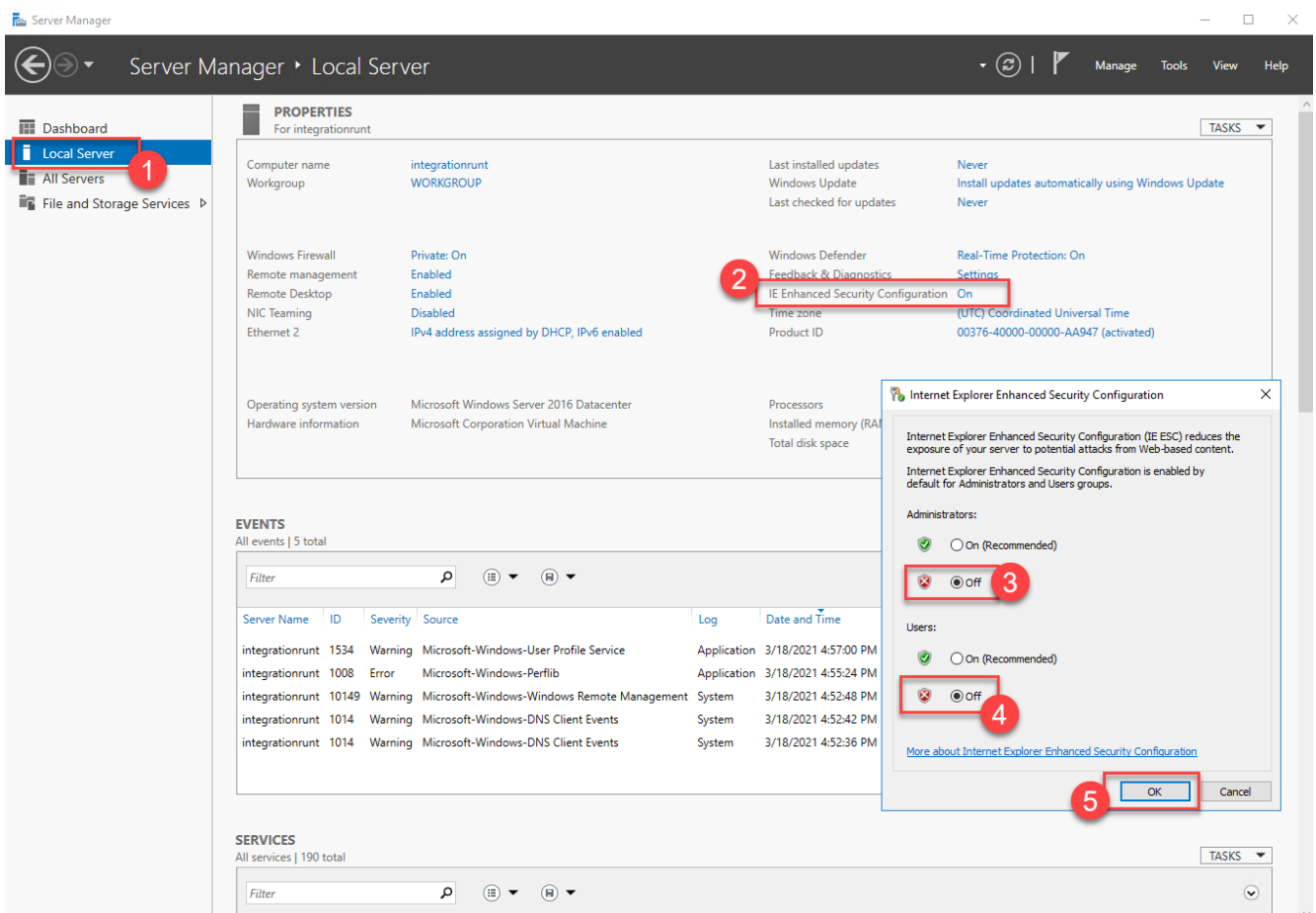
3389

**Download RDP File**

7. Open the RDP file. Enter the username and password you configured earlier. Disregard any certificate issues that RDP presents.



8. When you access the VM, **Server Manager** should open automatically. If not, open it manually using the search bar. Then, locate **Local Server** (1). Select **IE Enhanced Security Configuration**. Then, disable this feature for Administrators.



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