

Exercise 1: Create a Spark Cluster on HDInsight

Error when selecting 4 worker node:

[Home](#) > [Create a resource](#) >

Create HDInsight cluster ...

Basics Storage Security + networking **✖ Configuration + pricing** Tags Review + create

Configure cluster performance and pricing. [Learn More](#)

Node configuration

Configure your cluster's size and performance, and view estimated cost information.

The cost estimate represented in the table does not include subscription discounts or costs related to storage, networking, or data transfer.

✖ There are not enough cores available to support the selected number of nodes. Please adjust the number of nodes selected, pick a different region, or open a support case to request additional HDInsight cores.
[View cores usage](#)
[Open an HDInsight quota increase support case](#)

+ Add application

Node type	Node size	Number of ...	Estimated cost/h...
Head node	E8 V3 (8 Cores, 64 GB RAM), 0.71 USD/hour	2	1.42 USD
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.13 USD/hour	3	0.00 (FREE)
Worker node	E8 V3 (8 Cores, 64 GB RAM), 0.71 USD/hour	4	2.85 USD

☐ Enable autoscale
[Learn More](#)

Total estimated cost/hour 4.27 USD

Script actions

Use script actions to run custom PowerShell or Bash scripts on cluster nodes during cluster provisioning. [Learn about script actions](#)

Create HDInsight cluster ...

- Basics
- Storage
- Security + networking
- Configuration + pricing
- Tags
- Review + create

Configure cluster performance and pricing. [Learn More](#)

Node configuration

Configure your cluster's size and performance, and view estimated cost information.

The cost estimate represented in the table does not include subscription discounts or costs related to storage, networking, or data transfer.

This configuration will use 11 of 12 available cores in the West US 3 region.

[View cores usage](#)

[Open an HDInsight quota increase support case](#)

+ Add application

Node type	Node size	Number of ...	Estimated cost/h...
Head node	E2 V3 (2 Cores, 16 GB RAM), 0.18 USD/hour	2	0.36 USD
Zookeeper node	A1 v2 (1 Cores, 2 GB RAM), 0.06 USD/hour	3	0.18 USD
Worker node	A2m v2 (2 Cores, 16 GB RAM), 0.18 USD/hour	2	0.37 USD

☐ Enable autoscale

[Learn More](#)

Total estimated cost/hour 0.90 USD

Script actions

Use script actions to run custom PowerShell or Bash scripts on cluster nodes during cluster provisioning. [Learn about script actions](#)

+ Add script action

Create HDInsight cluster ...

Basics **Storage** Security + networking Configuration + pricing Tags Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type *

Azure Storage

Selection method * ⓘ

☒ Select from list

☐ Use access key

Primary storage account *

(New) sparklabhdistorage

Create new

Container * ⓘ

sparklab-2022-05-20t04-58-00-930z

Data Lake Storage Gen1

Provide details for the cluster to access Data Lake Storage Gen1. The cluster will be able to access any Data Lake Storage Gen1 accounts that the chosen service principal has access to.

Data Lake Storage Gen1 access [Configure access settings](#)

Additional Azure Storage

Link additional Azure Storage accounts to the cluster.

Account name

[Add Azure Storage](#)

Custom Ambari DB

Use an external Ambari database for greater flexibility, control, and customization. [Learn More](#)

Create HDInsight cluster ...

✔ Validation succeeded.

- Basics
- Storage
- Security + networking
- Configuration + pricing
- Tags
- Review + create

Spark 3.1 (HDI 4.0)

0.90 USD Total estimated cost/hour
This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.

Basics

Subscription	Azure subscription 1
Resource group	(new) spark-on-hdinsight
Region	West US 3
Cluster name	(new) sparklab
Cluster type	Spark 3.1 (HDI 4.0)
Cluster login username	admin
Secure Shell (SSH) username	sshuser
Use cluster login password for SSH	Enabled

Security + networking

Minimum TLS version	1.2
Resource provider connection	Inbound
Encryption at rest	Disabled
Encryption in transit	Disabled
Encryption at host on temp data disk	Disabled

Storage

Primary storage type	Azure Storage
Primary storage account	(new) hddisk
Container	sparklab-2022-05-20t04-58-00-930z

Create

« Previous

Next

Download a template for automation

Home >

spark-on-hdinsight Resource group

Search (Ctrl+F) << + Create Manage view Delete resource group Refresh Export to CSV Open query Assign tags Move Delete Export template Open in mobile

Overview

- Activity log
- Access control (IAM)
- Tags
- Resource visualizer
- Events

Settings

- Deployments
- Security
- Policies
- Properties
- Locks
- Cost Management

Essentials

Subscription (move) : [Azure subscription 1](#) Deployments : [1 Succeeded](#)

Subscription ID : 88045111-50ef-4121-b158-f449e5b40272 Location : West US 3

Tags (edit) : [Click here to add tags](#)

Resources Recommendations

Filter for any field... Type == all Location == all Add filter

Showing 1 to 2 of 2 records. Show hidden types

Name ↑↓	Type ↑↓	Location ↑↓
hdisk	Storage account	West US 3
sparklab	HDInsight cluster	West US 3

Exercise 2: Upload Jupyter Notebook to the cluster

Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER

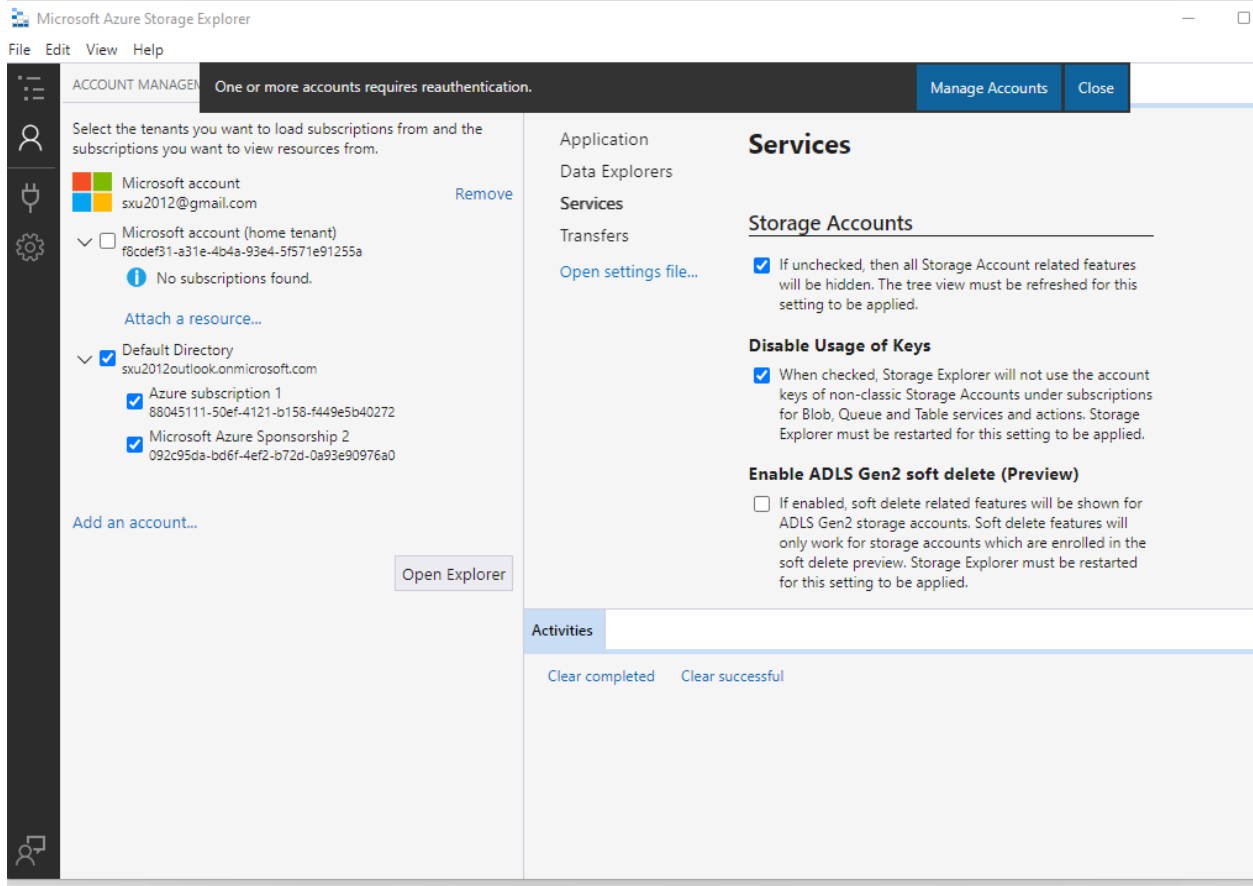
Search for resources

Collapse All Refresh All

- Quick Access
- Local & Attached
 - Storage Accounts
 - (Attached Containers)
 - (Emulator - Default Ports)
 - adfxudevstorage
 - Data Lake Storage Gen1 (Preview)

Activities

Clear completed Clear successful



Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER

This request is not authorized to perform this operation using this permission. RequestId:c4fd33c6-f01e-00b4-3e17-6c163f000000 Time:2022-05-10T10:37:10.37Z Close

Search for resources

Upload Download Open New Folder Select All Copy Paste Clone Delete Undelete Promote Version More

Local & Attached

- Storage Accounts
- Data Lake Storage Gen1 (Preview)
- Azure subscription 1 (xsu2012@gmail.com)
- Storage Accounts
 - hddisk
 - Blob Containers
 - sparklab-2022-05-20t04-58-00-930z
- Microsoft Azure Sponsorship 2 (xsu2012@gmail.com)
 - Storage Accounts
 - Disks
 - Data Lake Storage Gen1 (Preview)

File Shares

Queues

Tables

Disks

Data Lake Storage Gen1 (Preview)

Actions Properties

URL https://hddisk.blob.core.windows.net/

Custom Domain

Type Blob Container

HNS Enabled false

Lease State available

Lease Status unlocked

Public Read Access off

Last Modified Thursday, May 10, 2022 7:10:37 PM

Activities

Clear completed Clear successful

Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER

This request is not authorized to perform this operation using this permission. RequestId:95d0c3a-001e-008f-7717-6c539b000000 Time:2022-05-20T07:05:31.805Z Close

Search for resources

Upload Download Open New Folder Select All Copy Paste Clone Delete Undelete Promote Version Manage History More

Local & Attached

- Storage Accounts
- Data Lake Storage Gen1 (Preview)
- Azure subscription 1 (xsu2012@gmail.com)
- Storage Accounts
 - hddisk
 - Blob Containers
 - sparklab-2022-05-20t04-58-00-930z
- Microsoft Azure Sponsorship 2 (xsu2012@gmail.com)
 - Storage Accounts
 - Disks
 - Data Lake Storage Gen1 (Preview)

File Shares

Queues

Tables

Disks

Data Lake Storage Gen1 (Preview)

Actions Properties

URL https://hddisk.blob.core.windows.net/

Custom Domain

Type Blob Container

HNS Enabled false

Lease State available

Lease Status unlocked

Public Read Access off

Last Modified Thursday, May 10, 2022 7:10:37 PM

Activities

Clear completed Clear successful

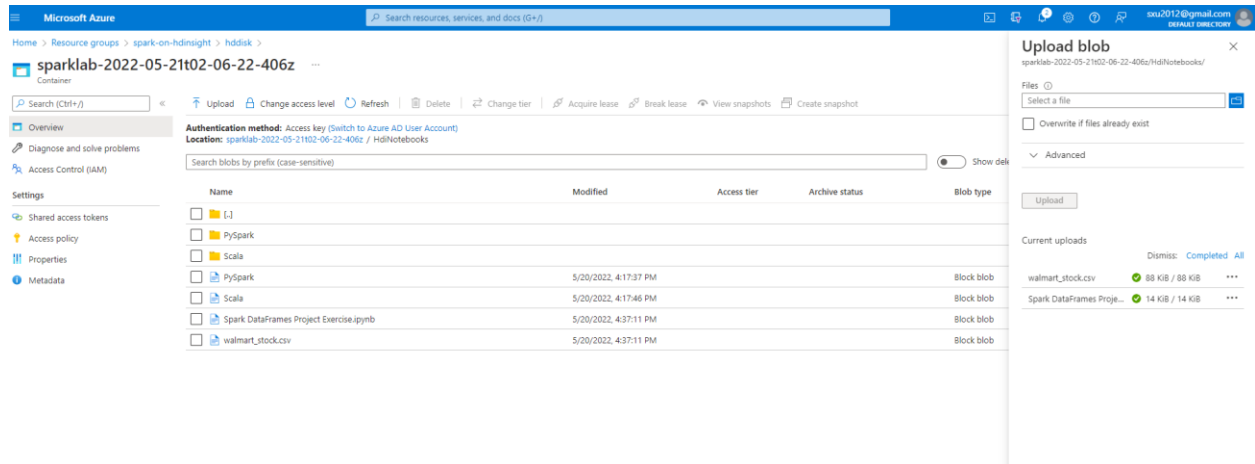
Transfer from 'C:\sb\projects\hdsight\...' to 'sparklab-2022-05-20t04-58-00-930z/HdiNotebooks/' failed: 0 items transferred, 1 item failed (used Azure AD, discovery completed)
Started at: 5/19/2022 9:09 PM, Duration: 3 seconds

Copy AzCopy Command to Clipboard Go to AzCopy Log File Retry...

Transfer from 'C:\sb\projects\hdsight\...' to 'sparklab-2022-05-20t04-58-00-930z/HdiNotebooks/' failed: 0 items transferred, 2 items failed (used Azure AD, discovery completed)
Started at: 5/19/2022 9:06 PM, Duration: 3 seconds

Copy AzCopy Command to Clipboard Go to AzCopy Log File Retry...

My Microsoft account seems to have some trouble, according to my mentor. So I uploaded the files using the web interface on Azure instead of the Storage Explorer



Microsoft Azure

Home > Resource groups > spark-on-hdinsight > hddisk > sparklab-2022-05-21t02-06-22-406z

Authentication method: Access key (Switch to Azure AD User Account)
Location: sparklab-2022-05-21t02-06-22-406z / HdINotebooks

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> [.]				
<input type="checkbox"/> PySpark				
<input type="checkbox"/> Scala				
<input type="checkbox"/> PySpark	5/20/2022, 4:17:37 PM			Block blob
<input type="checkbox"/> Scala	5/20/2022, 4:17:46 PM			Block blob
<input type="checkbox"/> Spark DataFrames Project Exercise.ipynb	5/20/2022, 4:37:11 PM			Block blob
<input type="checkbox"/> walmart_stock.csv	5/20/2022, 4:37:11 PM			Block blob

Upload blob

Files

Select a file

Overwrite if files already exist

Advanced

Upload

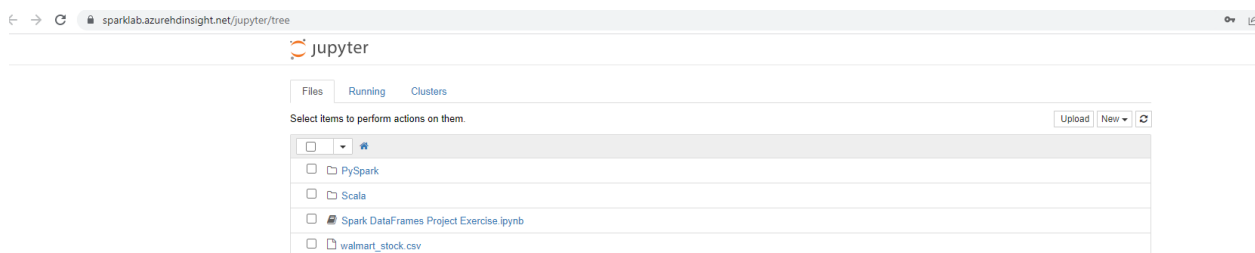
Current uploads

Dismiss Completed All

walmart_stock.csv 88 KiB / 88 KiB

Spark DataFrames Project Exercise.ipynb 14 KiB / 14 KiB

Exercise 3: Work with Jupyter Notebooks



sparklab.azurehdinsight.net/jupyter/tree

Jupyter

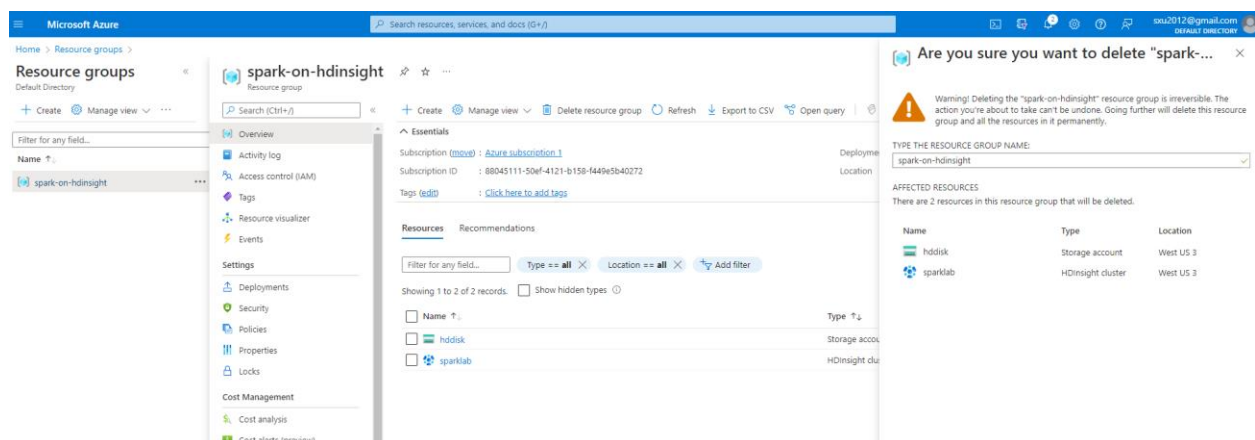
Files Running Clusters

Select items to perform actions on them.

Upload New

<input type="checkbox"/> [.]
<input type="checkbox"/> PySpark
<input type="checkbox"/> Scala
<input type="checkbox"/> Spark DataFrames Project Exercise.ipynb
<input type="checkbox"/> walmart_stock.csv

Exercise 4: Remove the HDInsight Spark cluster



Microsoft Azure

Home > Resource groups > spark-on-hdinsight

Resource groups

Filter for any field...

Name ↑

spark-on-hdinsight

Overview

Activity log

Access control (IAM)

Tags

Resource visualizer

Events

Settings

Deployments

Security

Policies

Properties

Locks

Cost Management

Cost analysis

Cost alerts (preview)

spark-on-hdinsight

Subscription (move) : Azure subscription 1

Subscription ID : 88045111-506f-4121-b158-449e5b40272

Tags (edit) : Click here to add tags

Resources

Filter for any field...

Type == all

Location == all

Add filter

Showing 1 to 2 of 2 records

Show hidden types

Name	Type	Location
hddisk	Storage account	West US 3
sparklab	HDInsight cluster	West US 3

Are you sure you want to delete "spark-..."

Warning! Deleting the "spark-on-hdinsight" resource group is irreversible. The action you're about to take can't be undone. Going further will delete this resource group and all the resources in it permanently.

TYPE THE RESOURCE GROUP NAME:

spark-on-hdinsight

AFFECTED RESOURCES

There are 2 resources in this resource group that will be deleted.

Name	Type	Location
hddisk	Storage account	West US 3
sparklab	HDInsight cluster	West US 3