Azure Mini Project: Spark on HDInsight

Estimated Time: 2-3 Hours

Overview

Today, data is being generated and stored in staggering amounts at unprecedented velocities. To make things more complicated, this data is generated in an ever-expanding variety of formats.

In order to gain actionable insights into Big Data sources, new tools need to be leveraged that allow the data to be cleaned, analyzed, and visualized quickly and efficiently. Azure HDInsight provides one solution to this problem. HDInsight makes it simple to create high-performance computing clusters equipped with Apache Spark and Spark ecosystem tools. This tool allows you to focus your time on the data itself rather than deploying, installing, configuring and maintaining the resources required to analyze the data.

Apache Spark is an open-source parallel-processing platform that excels at running large-scale data analytics jobs. Spark's combined use of in-memory and disk data storage delivers performance improvements that allow it to process some tasks up to 100 times faster than Hadoop. Azure makes deploying Spark clusters relatively simple and allows you to work on your data analysis sooner.

In this mini-project, you'll get hands-on experience with Apache Spark for Azure HDInsight. After provisioning a Spark cluster, you will use the Microsoft Azure Storage Explorer to add a Jupyter Notebook to the cluster. You will then use the notebook to explore and analyze Walmart stock data. Your goal is to learn how to create and utilize your own Spark clusters, provision them with Azure, and get a working introduction to Spark data analytics.

Learning Goals

In this mini-project, you will learn how to:

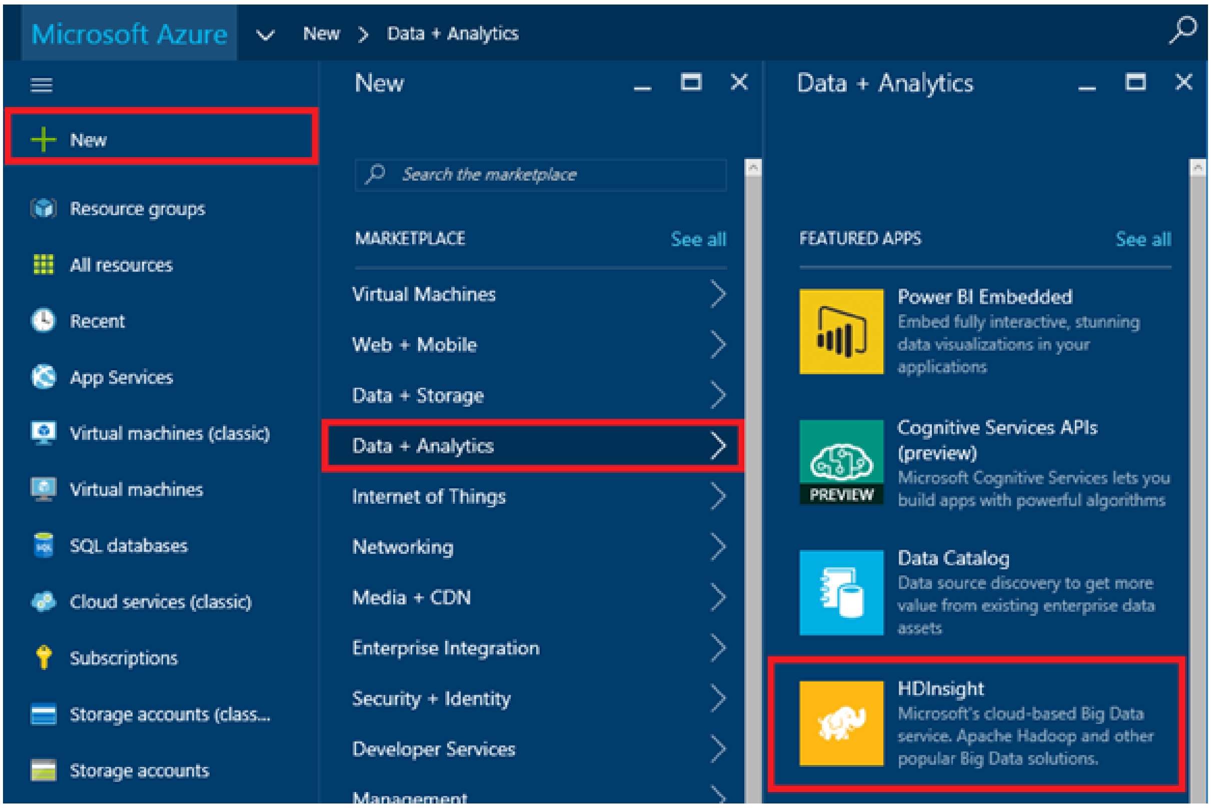
- Deploy an HDInsight Spark cluster
- Work with content stored in Azure Blob Storage and accessed by the Spark cluster as an HDFS volume
- Use a Jupyter Notebook to interactively explore a large dataset
- Delete a Spark cluster to avoid incurring unnecessary charges

Check the Deliverables at the end of this document before you begin working.

Exercise 1: Create a Spark Cluster on HDInsight

First create an HDInsight cluster running Apache Spark.

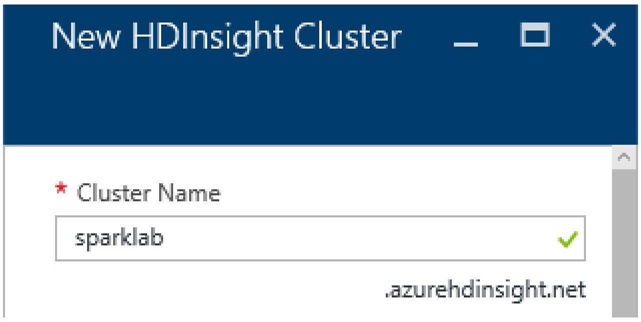
1. Go to the Azure Portal and sign in using your Microsoft account.
2. Click **+ New** in the upper-left corner. Then click **Data + Analytics**, followed by **HDInsight**.



Creating an HDInsight cluster

3. In the **Cluster Name** box, enter a unique DNS name for the cluster and make sure a green check mark appears next to it indicating that the name is valid and unique.

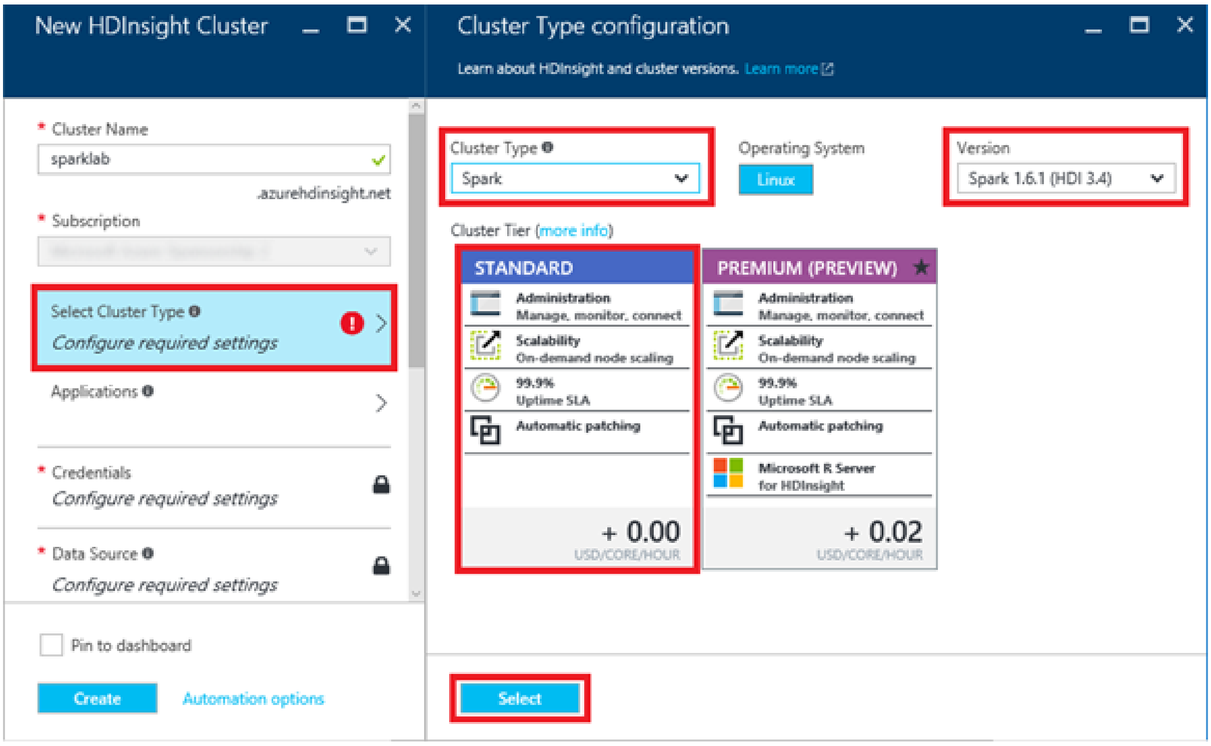
Make your cluster name as unique as possible by including birth dates, initials, and anything else you care to add to ensure no one else chooses the same name*. The name you entered may be unique right now, but it might NOT be unique a few minutes into the deployment.*



* Cluster Name: sparklab ✓ .azurehdinsight.net

Specifying the cluster name

4. Click **Select Cluster Type** to open a "Cluster Type configuration" blade. In that blade, select **Spark** as the **Cluster Type**. Under **Version**, select the latest Spark version, which is **1.6.1 (HDI 3.4)** at the time of this writing. Then click **Standard** to select a **Cluster Tier**, and finish up by clicking the **Select** button at the bottom of the blade.



* Cluster Name: sparklab ✓ .azurehdinsight.net

* Subscription

Cluster Type: Spark

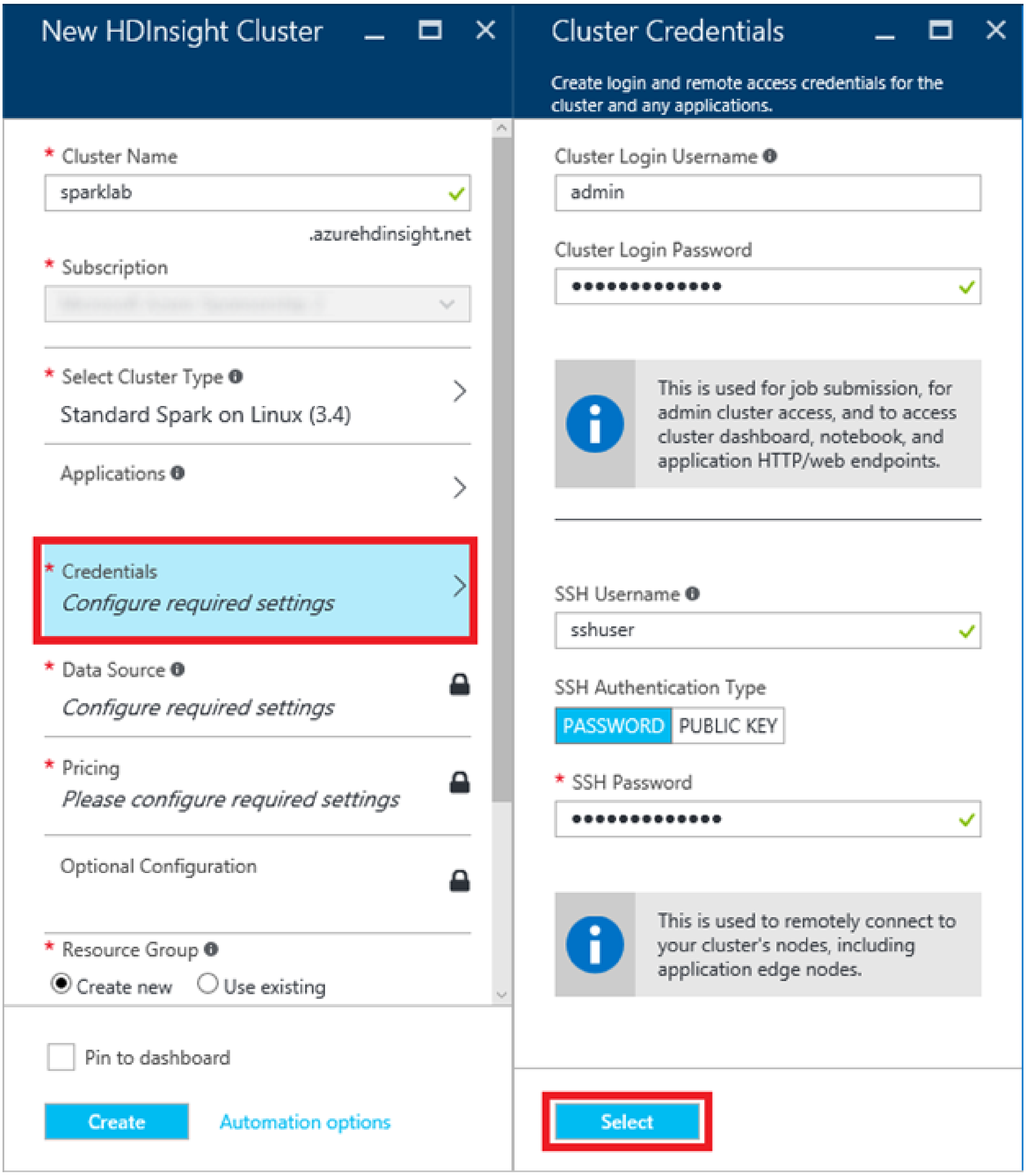
Operating System: Linux

Version: Spark 1.6.1 (HDI 3.4)

Specifying the cluster type

5. Click **Credentials** to open a "Cluster Credentials" blade. Leave **Cluster Login Username** set to "admin" and make sure **SSH Authentication Type** is set to **Password**. Enter "sshuser" (without quotation marks) for the **SSH Username** and "A4rsparkdemo!"

(again without quotation marks) for the **Cluster Login Password** and **SSH Password**. Then click the **Select** button at the bottom of the blade.



New HDInsight Cluster

* Cluster Name
sparklab ✓
.azurehdinsight.net

* Subscription

* Select Cluster Type
Standard Spark on Linux (3.4)

Applications

* Credentials
Configure required settings

* Data Source
Configure required settings

* Pricing
Please configure required settings

Optional Configuration

* Resource Group
● Create new ○ Use existing

□ Pin to dashboard

Create Automation options

Cluster Credentials

Create login and remote access credentials for the cluster and any applications.

Cluster Login Username
admin

Cluster Login Password
•••••••••••••• ✓

This is used for job submission, for admin cluster access, and to access cluster dashboard, notebook, and application HTTP/web endpoints.

SSH Username
sshuser ✓

SSH Authentication Type
PASSWORD PUBLIC KEY

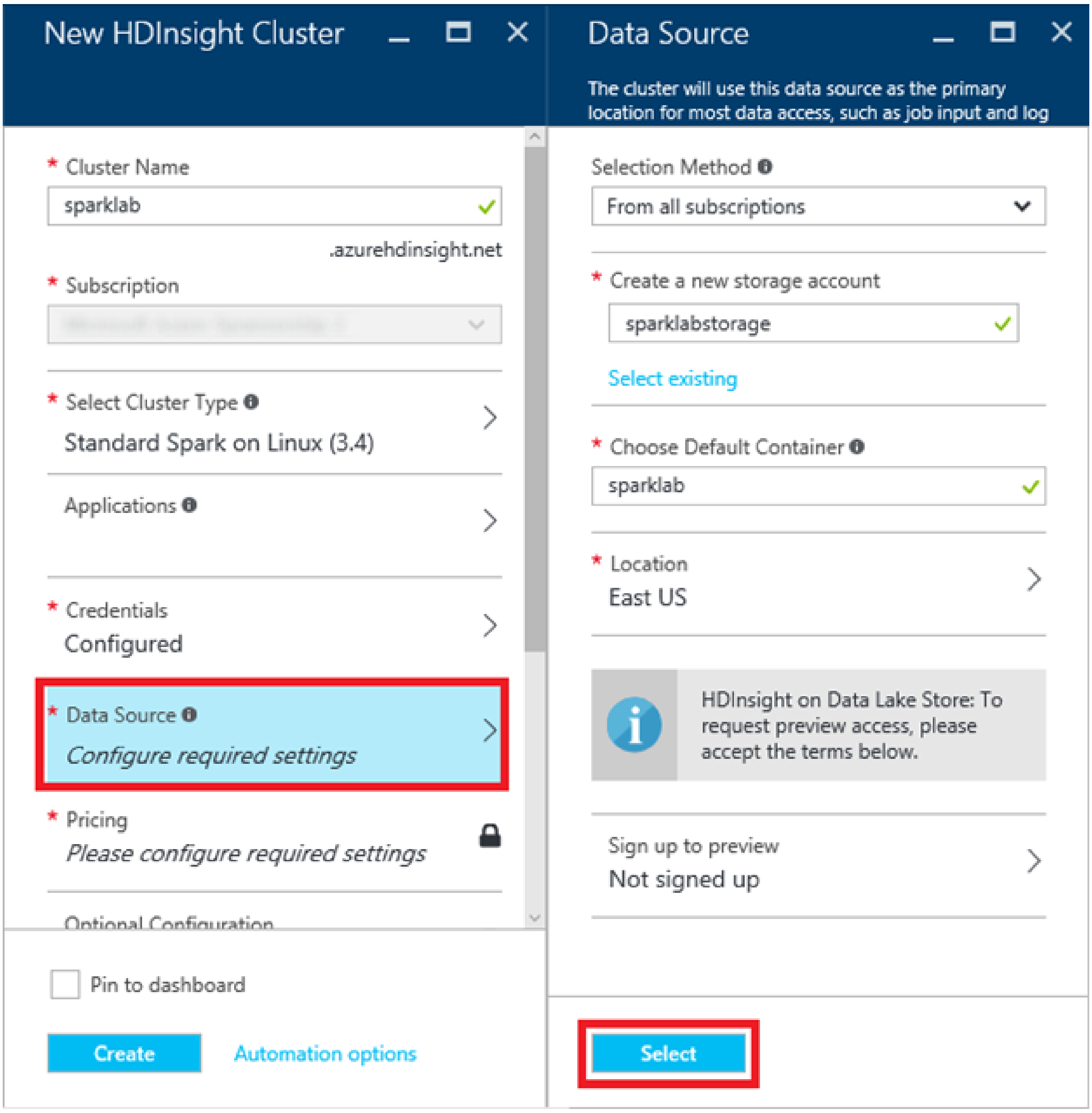
* SSH Password
•••••••••••••• ✓

This is used to remotely connect to your cluster's nodes, including application edge nodes.

Select

Specifying cluster credentials

6. Click **Data Source** to open a "Data Source" blade. Leave **Selection Method** set to **From all subscriptions** and enter a unique storage-account name in the box below **Create a new storage account**. Once more, make the name unique by including birth dates or other values unlikely to be used by someone else. For **Choose Default Container**, enter "sparklab" (without quotation marks). Select the **Location** nearest you, and then click the **Select** button at the bottom of the blade.



New HDInsight Cluster

* Cluster Name: sparklab ✓ .azurehdinsight.net

* Subscription: [redacted]

* Select Cluster Type: Standard Spark on Linux (3.4)

Applications

* Credentials: Configured

* Data Source: *Configure required settings*

* Pricing: *Please configure required settings*

Optional Configuration

□ Pin to dashboard

Data Source

The cluster will use this data source as the primary location for most data access, such as job input and log

Selection Method: From all subscriptions

* Create a new storage account: sparklabstorage ✓

* Choose Default Container: sparklab ✓

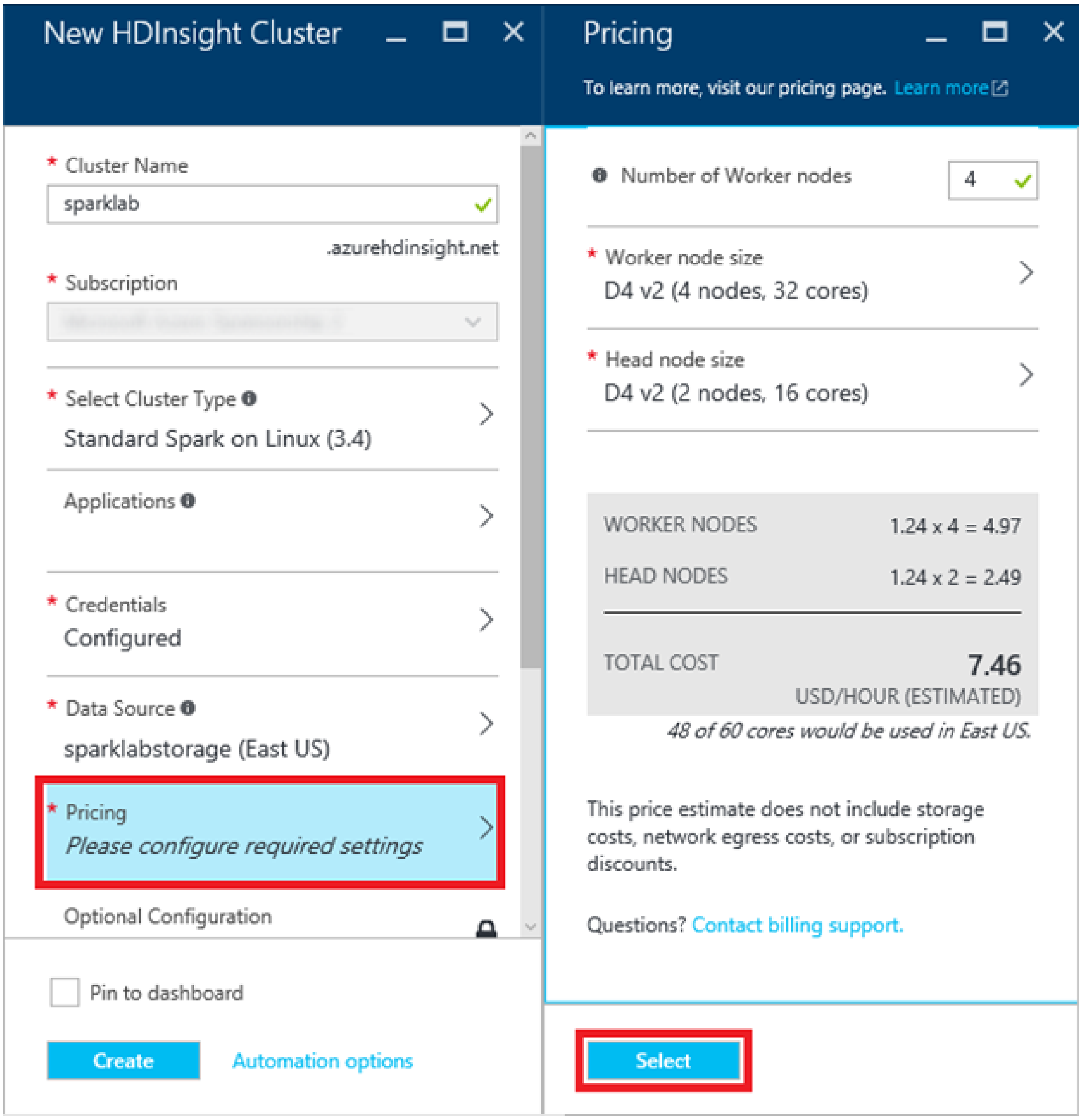
* Location: East US

HDInsight on Data Lake Store: To request preview access, please accept the terms below.

Sign up to preview: Not signed up

Specifying the data source

7. Click **Node Pricing Tiers** to open a "Pricing" blade. Make sure **Number of Worker nodes** is set to **4** and accept the default values everywhere else. Then click the **Select** button at the bottom of the blade.



New HDInsight Cluster

* Cluster Name
sparklab ✓
.azurehdinsight.net

* Subscription
[redacted]

* Select Cluster Type
Standard Spark on Linux (3.4)

Applications

* Credentials
Configured

* Data Source
sparklabstorage (East US)

* Pricing
Please configure required settings

Optional Configuration

□ Pin to dashboard

Pricing

To learn more, visit our pricing page. Learn more

Number of Worker nodes 4 ✓

* Worker node size
D4 v2 (4 nodes, 32 cores)

* Head node size
D4 v2 (2 nodes, 16 cores)

WORKER NODES	1.24 x 4 = 4.97
HEAD NODES	1.24 x 2 = 2.49
TOTAL COST	7.46 USD/HOUR (ESTIMATED)

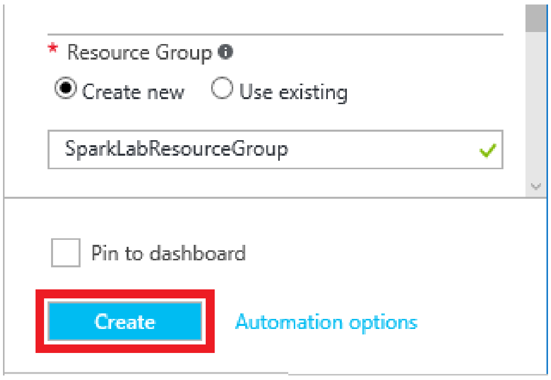
48 of 60 cores would be used in East US.

This price estimate does not include storage costs, network egress costs, or subscription discounts.

Questions? Contact billing support.

Specifying the node pricing tier

8. Select **Create new** under **Resource Group** and enter the resource-group name "SparkLabResourceGroup" (without quotation marks). Then click the **Create** button at the bottom of the blade to start deploying the cluster.



* Resource Group
● Create new ○ Use existing

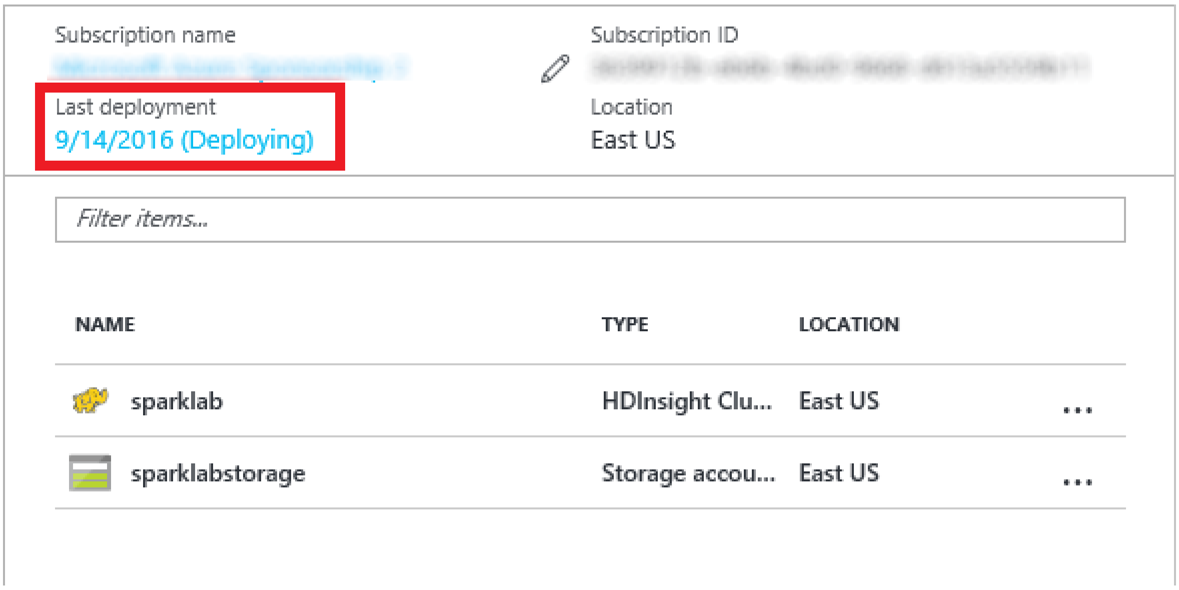
SparkLabResourceGroup ✓

□ Pin to dashboard

Specifying a resource group and creating the cluster

9. Deploying an HDInsight cluster can take 20 minutes or more. You can monitor the status of the deployment by opening the resource group's blade. Click **Resource group** in the ribbon on the left side of the portal, and then click the resource group name ("SparkLabResourceGroup") to open the blade. "Deploying" will change to "Succeeded" when the deployment has completed successfully.

*Click the browser's **Refresh** button every few minutes to update the deployment status. Clicking the **Refresh** button in the resource-group blade refreshes the list of resources in the resource group, but does not reliably update the deployment status.*



Subscription name
[redacted]

Subscription ID
[redacted]

Last deployment
9/14/2016 (Deploying)

Location
East US

Filter items...

NAME	TYPE	LOCATION	
sparklab	HDInsight Clu...	East US	...
sparklabstorage	Storage accou...	East US	...

Monitoring the deployment

Summary

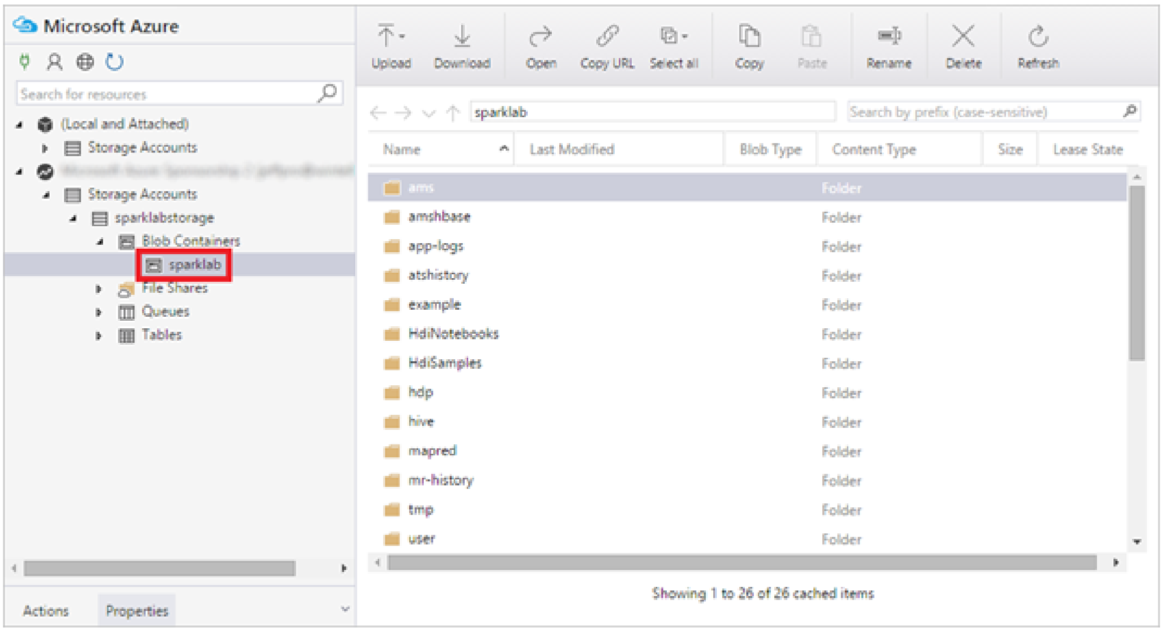
In this exercise, you learned how to provision an HDInsight Spark cluster on Azure and some of the available configurations you can choose from. Wait for the deployment to finish, and then proceed to the next exercise.

Exercise 2: Upload Jupyter Notebook to the cluster

You will be using Jupyter Notebooks to do the data-exploration and analysis portions of this lab. A notebook and data file have been prepared for you (**download them here: .ipynb file**, **data file**) and need to be uploaded to your cluster. In this exercise, you'll use the cross-platform Microsoft Azure Storage Explorer to upload the notebook. If Storage Explorer isn't installed on your computer, take the time to install it now.

1. Start the Microsoft Azure Storage Explorer. If you're prompted for credentials, sign in with the username and password for your Microsoft account.

2. In the Storage Explorer window, find the storage account you created when you deployed the Spark cluster (in Step 6 of the previous exercise). Expand the list of items underneath that storage account and click the small arrow next to **Blob Containers** to show a list of containers. Then click the container named "sparklab." It contains several folders created during the provisioning process. sparklab is the root folder for your Spark cluster.

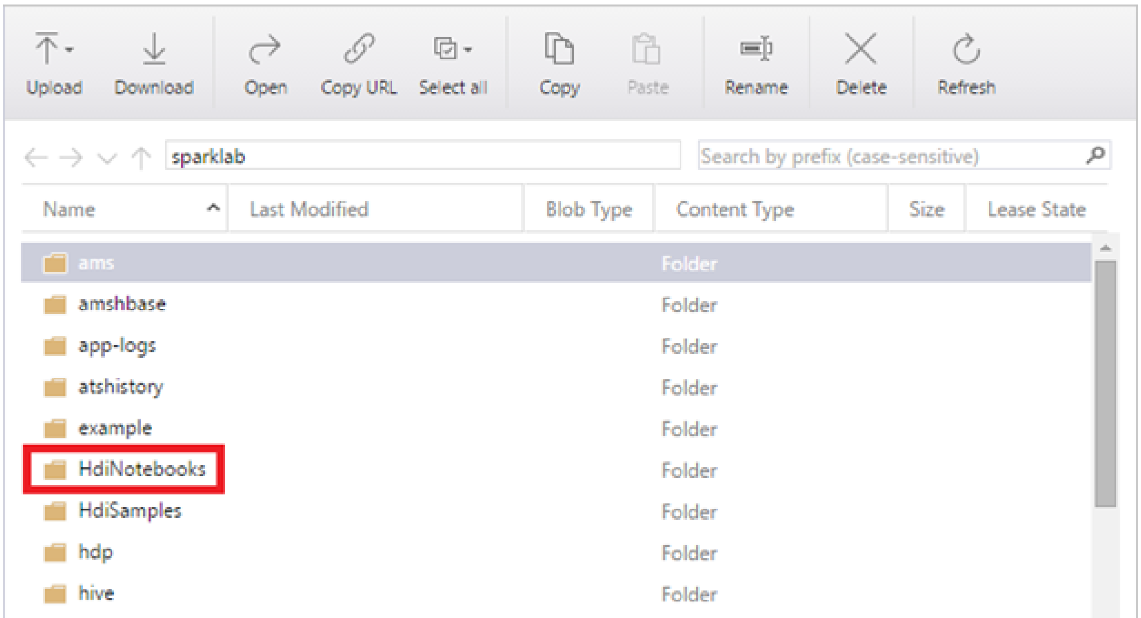


Name	Last Modified	Blob Type	Content Type	Size	Lease State
ams			Folder		
amshbase			Folder		
app-logs			Folder		
atshistory			Folder		
example			Folder		
HdiNotebooks			Folder		
HdiSamples			Folder		
hdp			Folder		
hive			Folder		
mapred			Folder		
mr-history			Folder		
tmp			Folder		
user			Folder		

Showing 1 to 26 of 26 cached items

Contents of the "sparklab" container

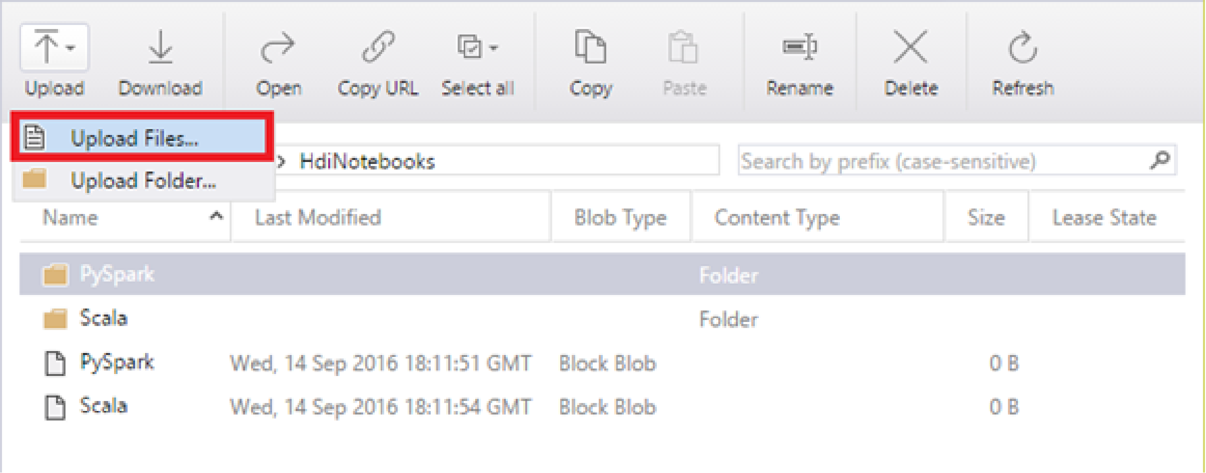
3. Double-click the folder named "HdiNotebooks." This is the root folder for the cluster's Jupyter notebooks.



Name	Last Modified	Blob Type	Content Type	Size	Lease State
ams			Folder		
amshbase			Folder		
app-logs			Folder		
atshistory			Folder		
example			Folder		
HdiNotebooks			Folder		
HdiSamples			Folder		
hdp			Folder		
hive			Folder		

Opening the "HdiNotebooks" folder

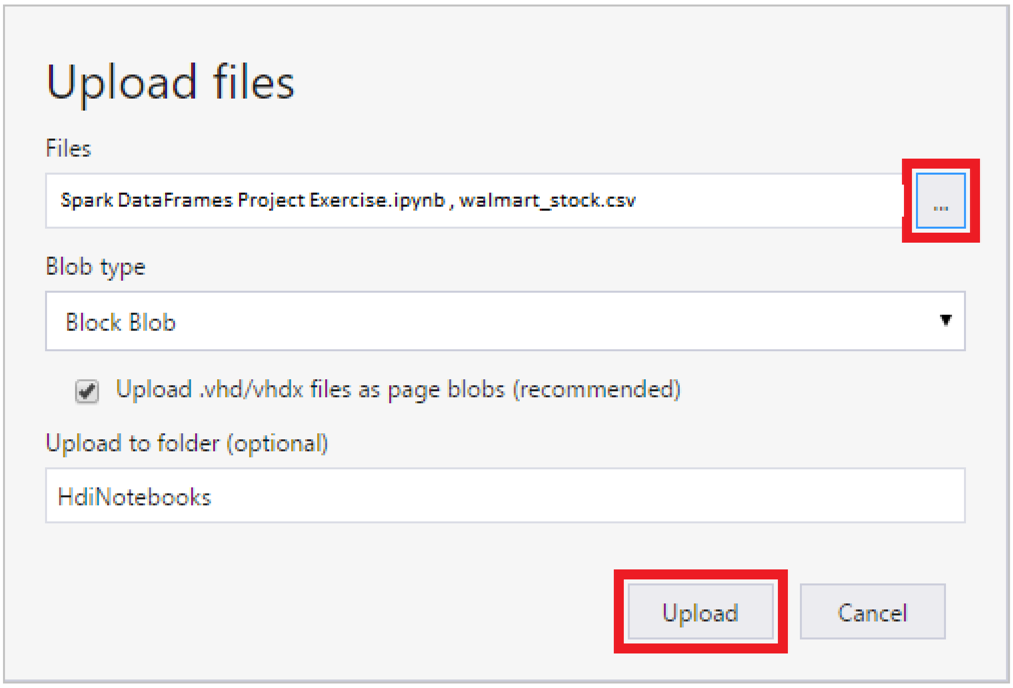
4. Click the **Upload** button and select **Upload Files** from the menu.



Name	Last Modified	Blob Type	Content Type	Size	Lease State
PySpark			Folder		
Scala			Folder		
PySpark	Wed, 14 Sep 2016 18:11:51 GMT	Block Blob		0 B	
Scala	Wed, 14 Sep 2016 18:11:54 GMT	Block Blob		0 B	

Uploading notebook to blob storage

5. Click the **...** button to the right of the field labeled "Files." In the ensuing dialog, navigate to this lab's "resources" subdirectory and select the .ipynb file and the walmart stock data file in that subdirectory. Then close the dialog and click the **Upload** button.



Files: Spark DataFrames Project Exercise.ipynb , walmart_stock.csv

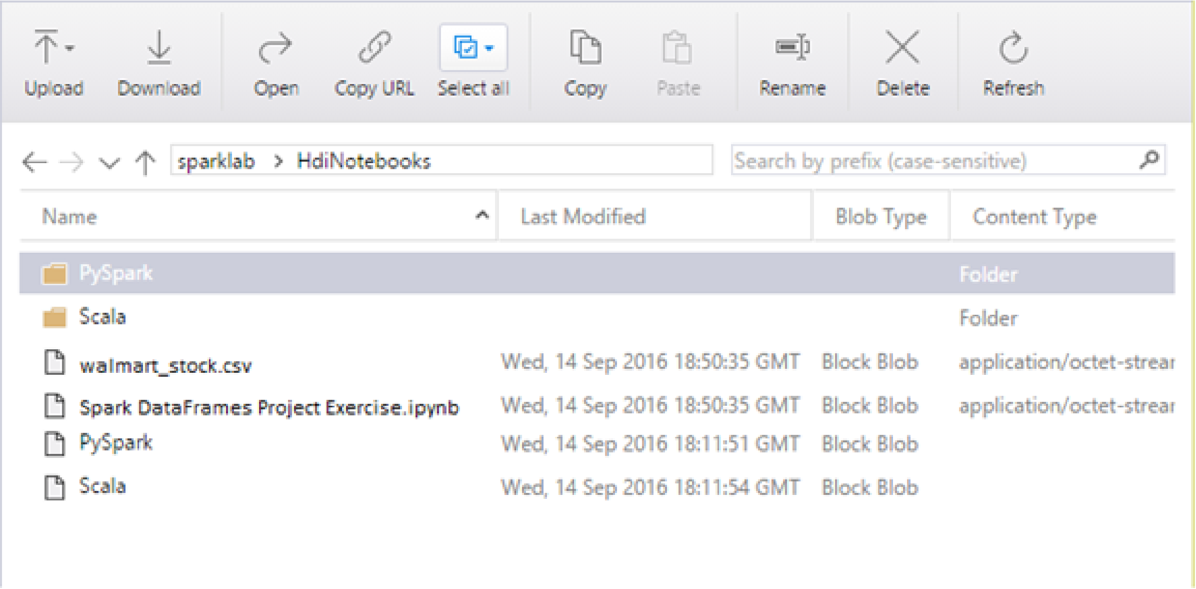
Blob type: Block Blob

☑ Upload .vhd/vhdx files as page blobs (recommended)

Upload to folder (optional): HdiNotebooks

Uploading .ipynb file and Walmart Stock Data File

6. Confirm that all three files were uploaded to the "HdiNotebooks" folder.



Name	Last Modified	Blob Type	Content Type
PySpark			Folder
Scala			Folder
walmart_stock.csv	Wed, 14 Sep 2016 18:50:35 GMT	Block Blob	application/octet-strear
Spark DataFrames Project Exercise.ipynb	Wed, 14 Sep 2016 18:50:35 GMT	Block Blob	application/octet-strear
PySpark	Wed, 14 Sep 2016 18:11:51 GMT	Block Blob	
Scala	Wed, 14 Sep 2016 18:11:54 GMT	Block Blob	

Uploaded notebook and walmart stock data

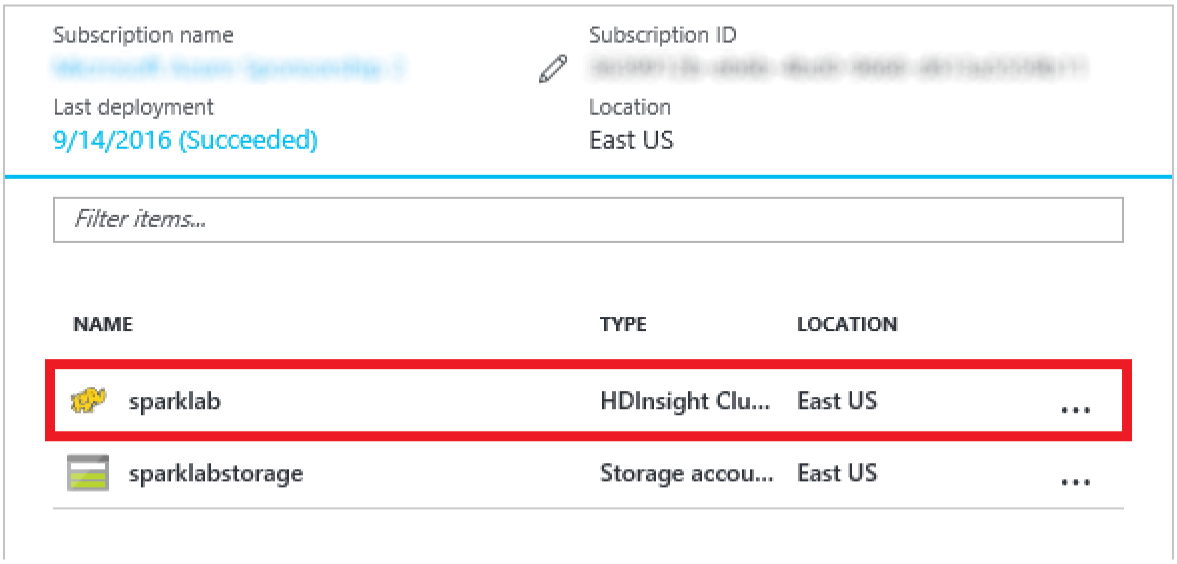
The notebook and data is uploaded and ready to go. Let's put them to work!

Exercise 3: Work with Jupyter Notebooks

Jupyter Notebooks are extraordinarily useful for data professionals. Jupyter supports several programming languages through the use of installable interpreters called *kernels*. Spark clusters on HDInsight include the Spark and PySpark kernels for Scala and Python, respectively.

In this exercise, you will learn how to access Jupyter notebooks in your Spark cluster and acquire basic skills for using them.

1. In the Azure Portal, return to the blade for the resource group ("SparkLabResourceGroup") that contains the cluster. In the list of resources that belong to the resource group, click the HDInsight cluster.



Subscription name
[redacted]

Subscription ID
[redacted]

Last deployment
9/14/2016 (Succeeded)

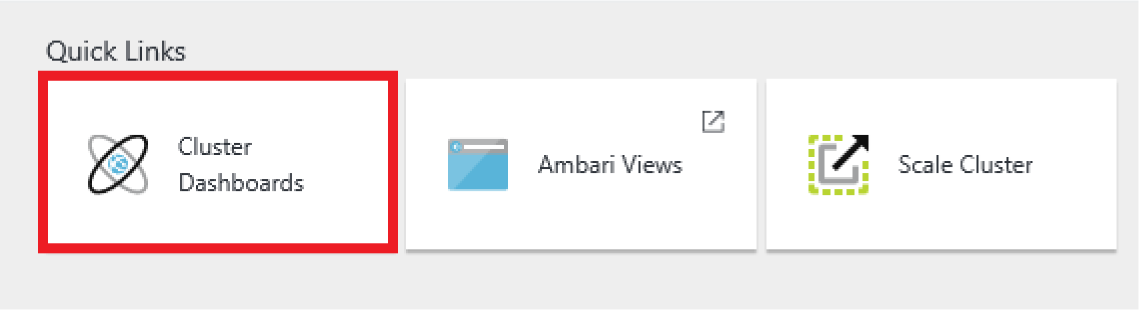
Location
East US

Filter items...

NAME	TYPE	LOCATION	
sparklab	HDInsight Clu...	East US	...
sparklabstorage	Storage accou...	East US	...

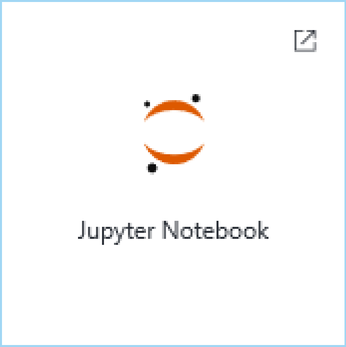
Opening the cluster

2. In the blade for the HDInsight Spark cluster, click **Cluster Dashboards.**



Opening the cluster dashboards

3. In the ensuing blade, click **Jupyter Notebook**.

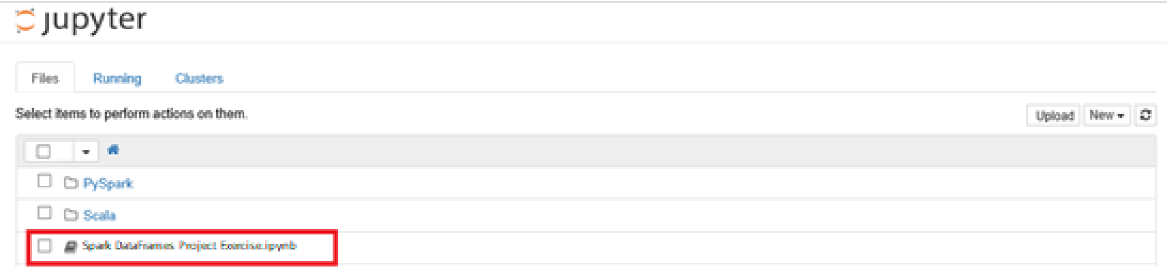


Opening a Jupyter notebook

4. When prompted for a username and password, log in with your cluster credentials ("admin" and "A4rsparkdemo!") from Exercise 1, Step 5.

If you mistype the password and are greeted with a 403 Forbidden error, start a new incognito or private browsing session, go to the Azure Portal, and open the Jupyter notebook again.

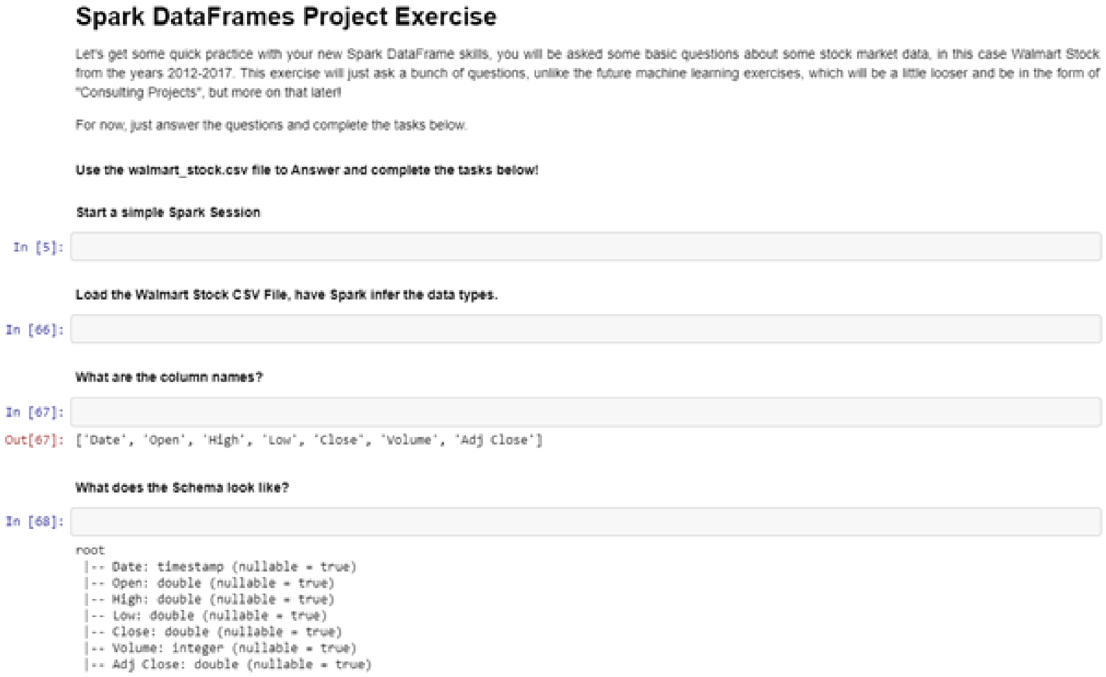
5. A new browser window (or tab) will open showing the Jupyter notebooks in your cluster. Here, you can manage your notebooks, upload new ones, and more. You can also see which notebooks are currently "running", meaning they are currently consuming resources in your Spark cluster. Confirm that you see the notebook you uploaded in Exercise 2. Then click **Spark Dataframes Project Exercises.ipynb** to open that notebook.



Select items to perform actions on them.

Opening Notebook

6. Jupyter notebooks consist of a series of cells into which you can insert commands, HTML, or Markdown text. The notebook you opened contains the remaining instructions for this exercise. Follow the instructions in the notebook to complete Exercise 3.



Spark DataFrames Project Exercise

Let's get some quick practice with your new Spark DataFrame skills, you will be asked some basic questions about some stock market data, in this case Walmart Stock from the years 2012-2017. This exercise will just ask a bunch of questions, unlike the future machine learning exercises, which will be a little looser and be in the form of "Consulting Projects", but more on that later!

For now, just answer the questions and complete the tasks below.

Use the walmart_stock.csv file to Answer and complete the tasks below!

Start a simple Spark Session

```
In [5]:
```

Load the Walmart Stock CSV File, have Spark infer the data types.

```
In [66]:
```

What are the column names?

```
In [67]:
Out[67]: ['Date', 'Open', 'High', 'Low', 'Close', 'Volume', 'Adj Close']
```

What does the Schema look like?

```
In [68]:

root
 |-- Date: timestamp (nullable = true)
 |-- Open: double (nullable = true)
 |-- High: double (nullable = true)
 |-- Low: double (nullable = true)
 |-- Close: double (nullable = true)
 |-- Volume: integer (nullable = true)
 |-- Adj Close: double (nullable = true)
```

Working with Jupyter notebook

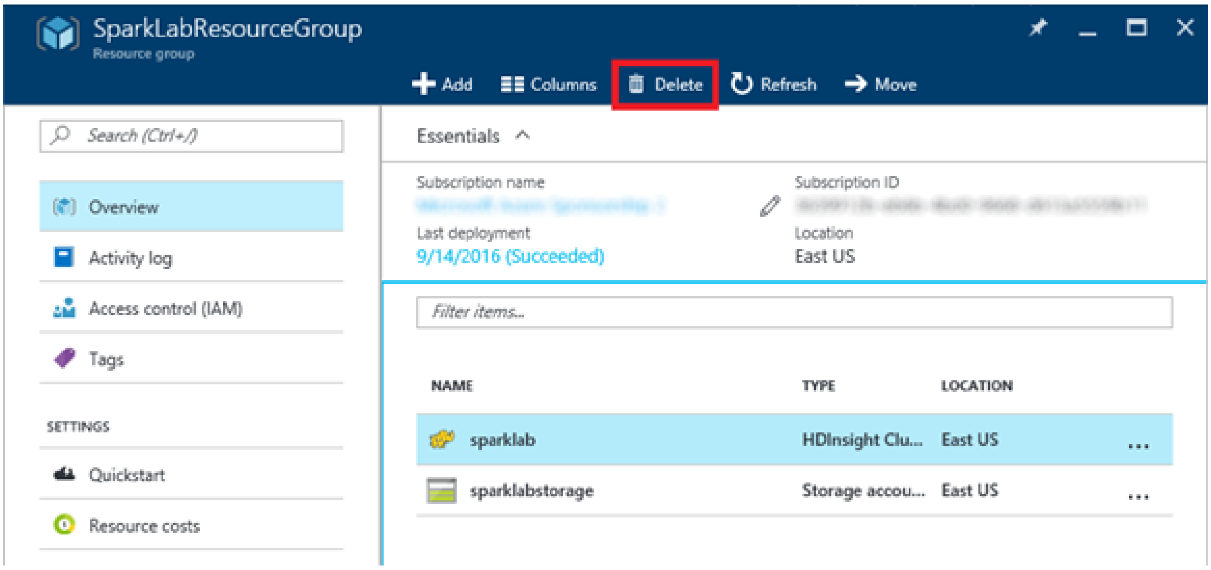
Once you have completed the steps in the notebook, save and checkpoint the notebook then close the browser window in which the notebook is displayed and return to the Azure Portal.

7. Download the **Spark Dataframes Project Exercises.ipynb** to your computer by right clicking the file in Azure Storage Explorer and clicking ‘Download’. You will submit the Jupyter Notebook to your mentor .

Exercise 4: Remove the HDInsight Spark cluster

When you are finished using an HDInsight Spark cluster, you should delete it. **You are charged for it while it exists, regardless of whether it's doing any work.** In this exercise, you will delete the resource group created in Exercise 1 when you created the cluster. Deleting the resource group deletes everything in it and prevents any further charges.

1. In the Azure Portal, open the blade for the "SparkLabResourceGroup" resource group that holds the cluster. Then click the **Delete** button at the top of the blade.



Subscription name
[redacted]

Subscription ID
[redacted]

Last deployment
9/14/2016 (Succeeded)

Location
East US

NAME	TYPE	LOCATION
sparklab	HDInsight Clu...	East US
sparklabstorage	Storage accou...	East US

Deleting a resource group

2. For safety, you are required to type in the resource group's name. **Once deleted, a resource group cannot be recovered.** Type the name of the resource group. Then click the **Delete** button to remove all traces of this lab from your account.

After a few minutes, the cluster and all of its resources will be deleted. Billing stops when you click the **Delete** button, so you're not charged for the time required to delete the cluster. Similarly, billing doesn't start until a cluster is fully and successfully deployed.

Summary

Here is a summary of what you learned in this lab:

- Apache Spark for Azure HDInsight is Microsoft Azure's implementation of Hadoop, Spark, and several other Big Data tools
- The Azure Portal makes it easy to create, configure, and delete HDInsight Spark clusters
- HDInsight Spark clusters come with Jupyter preinstalled
- Jupyter Notebooks provide a powerful means for querying and analyzing data
- HDInsight Spark clusters should be deleted when they're no longer needed to avoid incurring unwanted charges

With Apache Spark for Azure HDInsight, high-performance computing clusters with all the tools you need to handle big data are just a few button clicks away. It's just one example of why cloud computing is changing the face of research.

Deliverables

1. Jupyter Notebook with code to achieve stated objective in each cell
2. A PPT/Word-doc with screenshots of each step as performed according to this lab