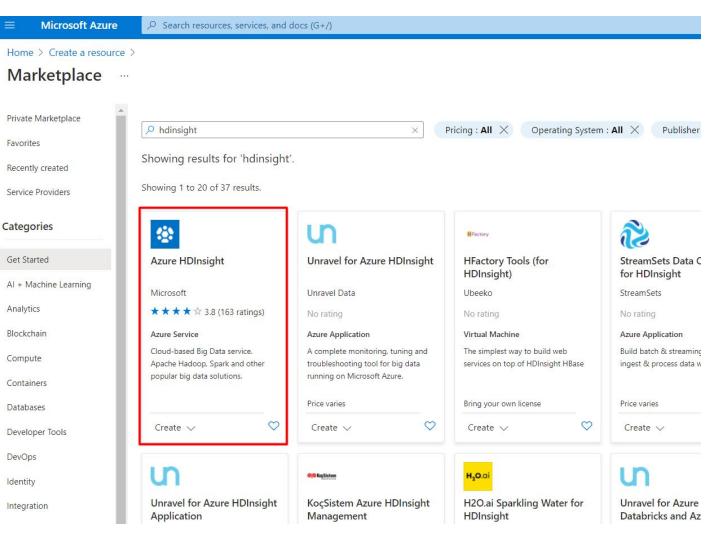
After logging into your Azure Portal, click "Create a resource". This will take you to the Marketplace where you can search for Azure HDInsight.



Fill in the required fields and make sure to select Spark as the Cluster type and pick the latest version.

Create HDInsight cluster

Create a managed HDInsight cluster. Select from Spark, Kafka, Hadoop, Storm, and more. 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.

ubscription *	Azure subscription 1	~
Resource group *	(New) rg-sb-free-01	~
	Create new	×

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. Learn more

Cluster name *	sb-sparklab-042121	~
Region *	West US	V
Cluster type *	Spark	
1	Change	
Version *	Spark 2.4 (HDI 4.0)	~

Cluster credentials

Use cluster login password for SSH

Enter new credentials that will be used to administer or access the cluster.

Cluster login username * ①	admin	
Cluster login password *		~
Confirm cluster login password *		~
Secure Shell (SSH) username * ①	sshuser	
	_	

Create HDInsight cluster

Configuration + pricing Security + networking Tags Review + create Basics 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 Select from list () Use access key Selection method * (i) Primary storage account * (New) sbsparklab042hdistorage V Create new sb-sparklab-042121-2021-04-21t18-50-37-763z Container * (1) 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.

Add Azure Storage

Custom Ambari DB

Use an external Ambari database for greater flexibility, control, and customization. Learn More

Since there is a maximum of 40 available cores, I tweaked the number of worker nodes to fit the limit. The data we are processing wasn't large so there weren't any issues. After configuring the cluster size, hit the create button and wait for completion.

Create HDInsight cluster

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.



Node size

+ Add application Node type

 Head node
 E8 V3 (8 Cores, 64 GB RAM), 0.66 USD/hour
 ✓
 2
 1.31 USD

 Zookeeper node
 A2 v2 (2 Cores, 4 GB RAM), 0.13 USD/hour
 ✓
 3
 0.00 (FREE)

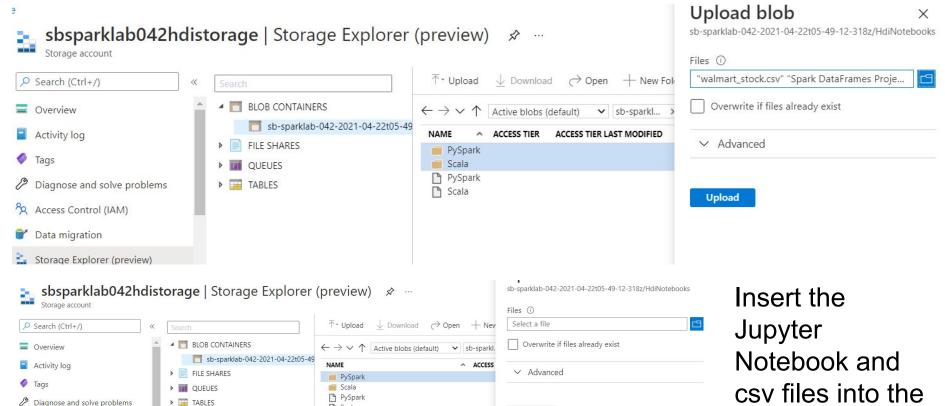
 Worker node
 E8 V3 (8 Cores, 64 GB RAM), 0.66 USD/hour
 ✓
 2
 ✓
 1.31 USD

Number of ...

Estimated cost/h...

Enable autoscale

Total estimated cost/hour 2.62 USD



Spark DataFrames Project Exercise.ipvnb

walmart_stock.csv

Access Control (IAM)

Storage Explorer (preview)

Data migration

Access keys

Settings

Upload

Current uploads

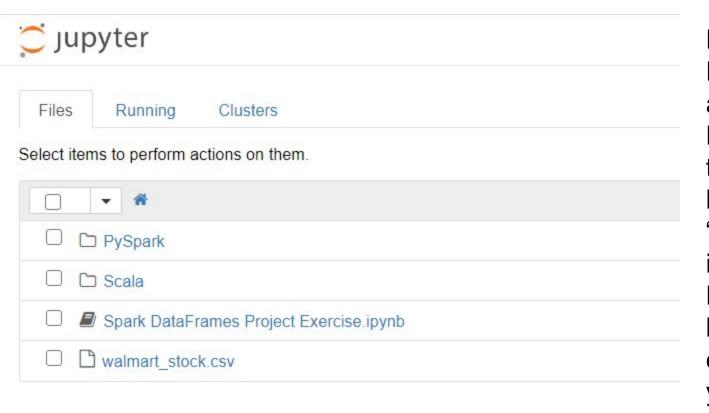
walmart stock.csv

Spark DataFrames Proie... 4 KiB / 14 KiB

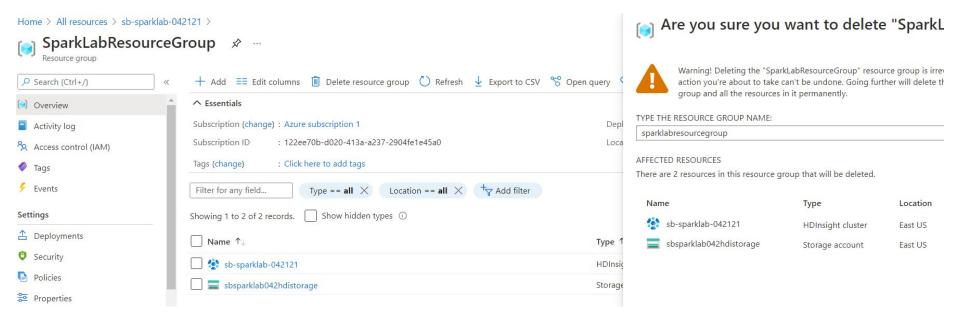
88 KiB / 88 KiB

Dismiss: Completed All

csv files into the "HdiNotebooks" folder.



Return to the **HDInsight Cluster** and find Jupyter Notebook under the dashboards header in "Overview". This is launch Jupyter Notebook in your browser where can see the files you uploaded earlier.



Be sure to delete the cluster once the tasks are completed to avoid unnecessary charges.