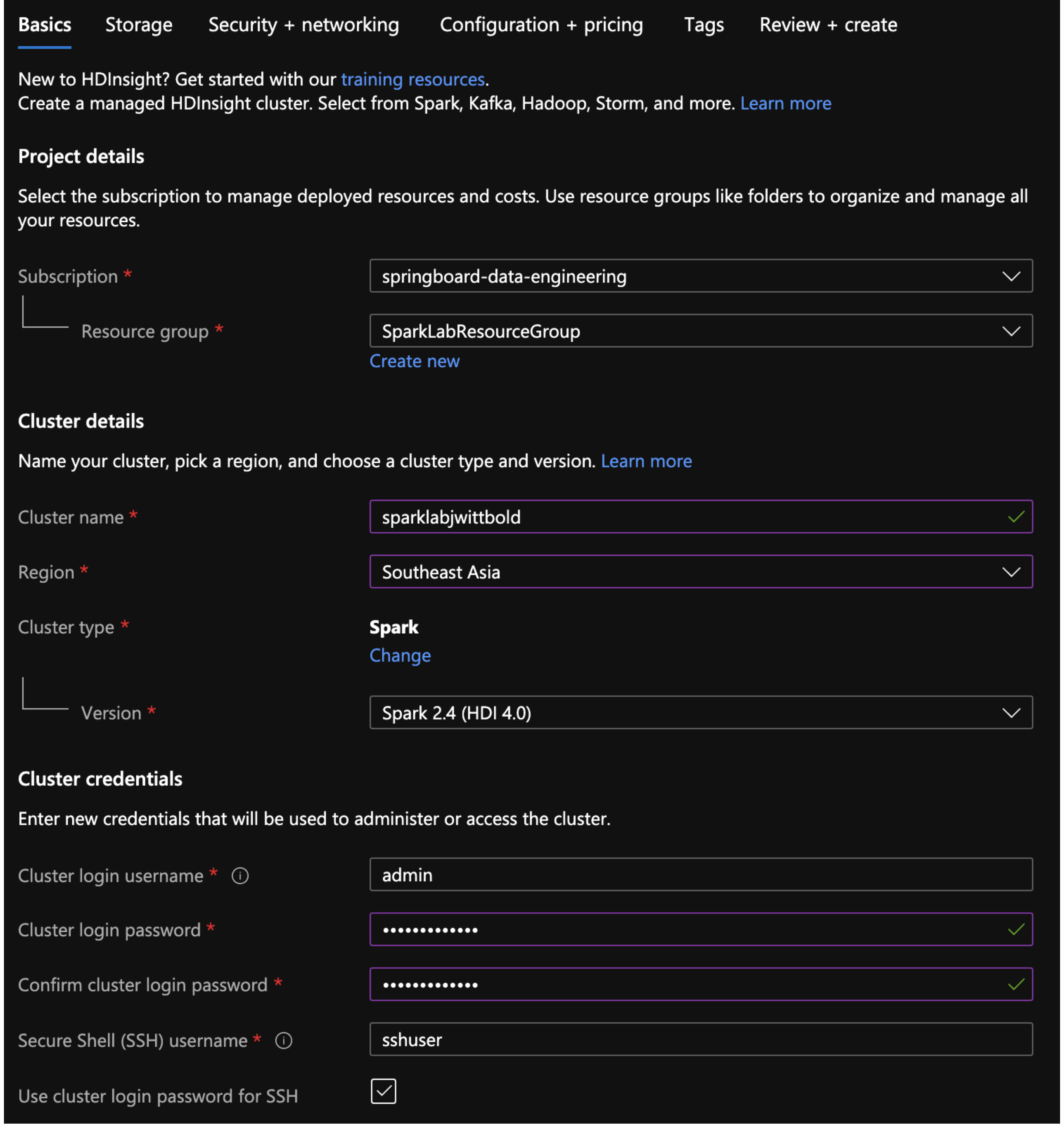
Deploying HDInsight Spark Cluster on Azure

Step by Step



New to HDInsight? Get started with our training resources.
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.

Subscription *	springboard-data-engineering
Resource group *	SparkLabResourceGroup Create new

Cluster details

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

Cluster name *	sparklabjwittbold ✓
Region *	Southeast Asia
Cluster type *	**Spark** Change
Version *	Spark 2.4 (HDI 4.0)

Cluster credentials

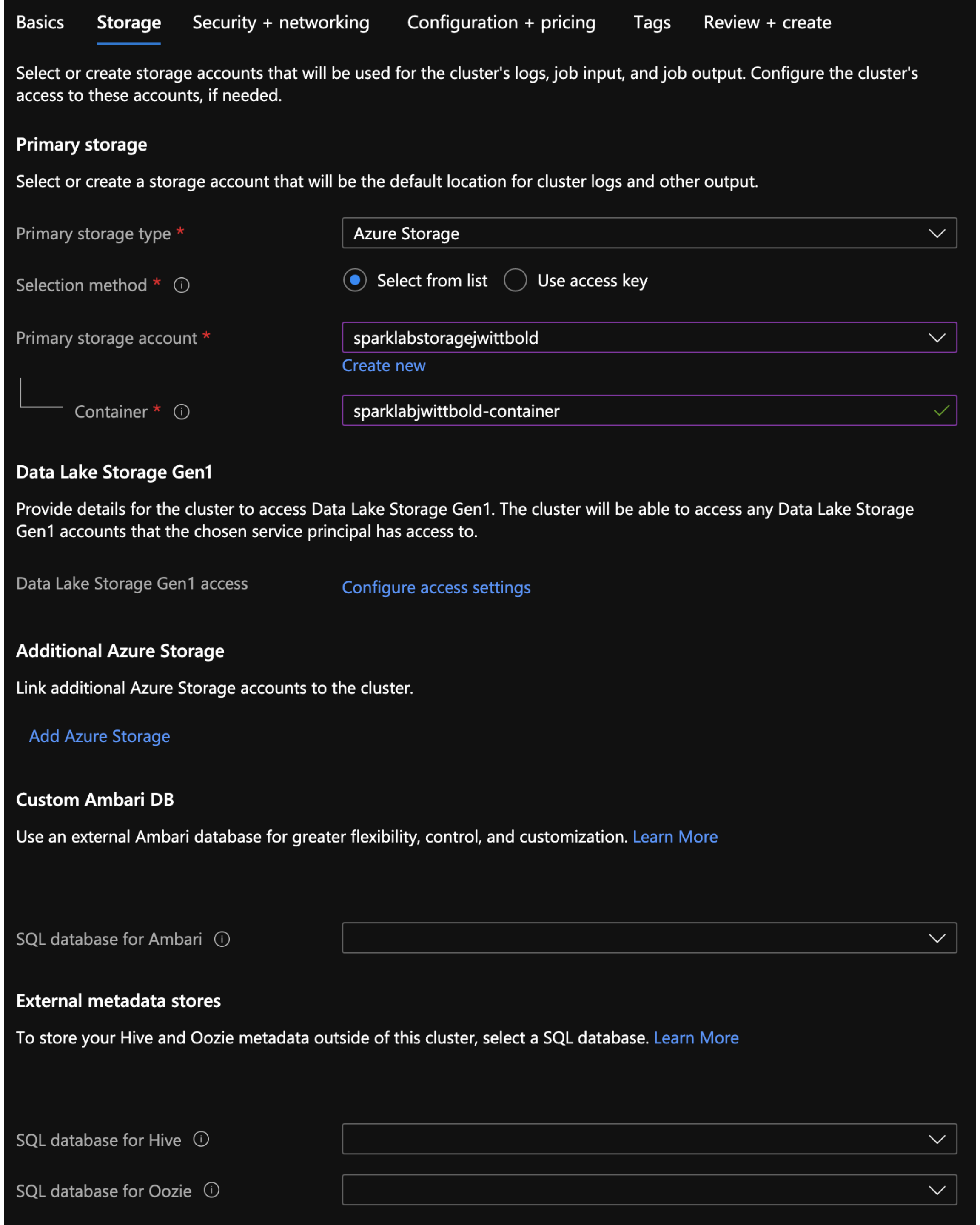
Enter new credentials that will be used to administer or access the cluster.

Cluster login username * ⓘ	admin
Cluster login password *	••••••••••••• ✓
Confirm cluster login password *	••••••••••••• ✓
Secure Shell (SSH) username * ⓘ	sshuser
Use cluster login password for SSH	☑

- Select Subscription
- Select Resource Group

- Provide Cluster name
- Choose Region
- Specify Cluster Type / Version

- Set password to login as admin over SSH



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 * sparklabstoragejwittbold

Create new

Container * sparklabjwittbold-container ✓

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

SQL database for Ambari

External metadata stores

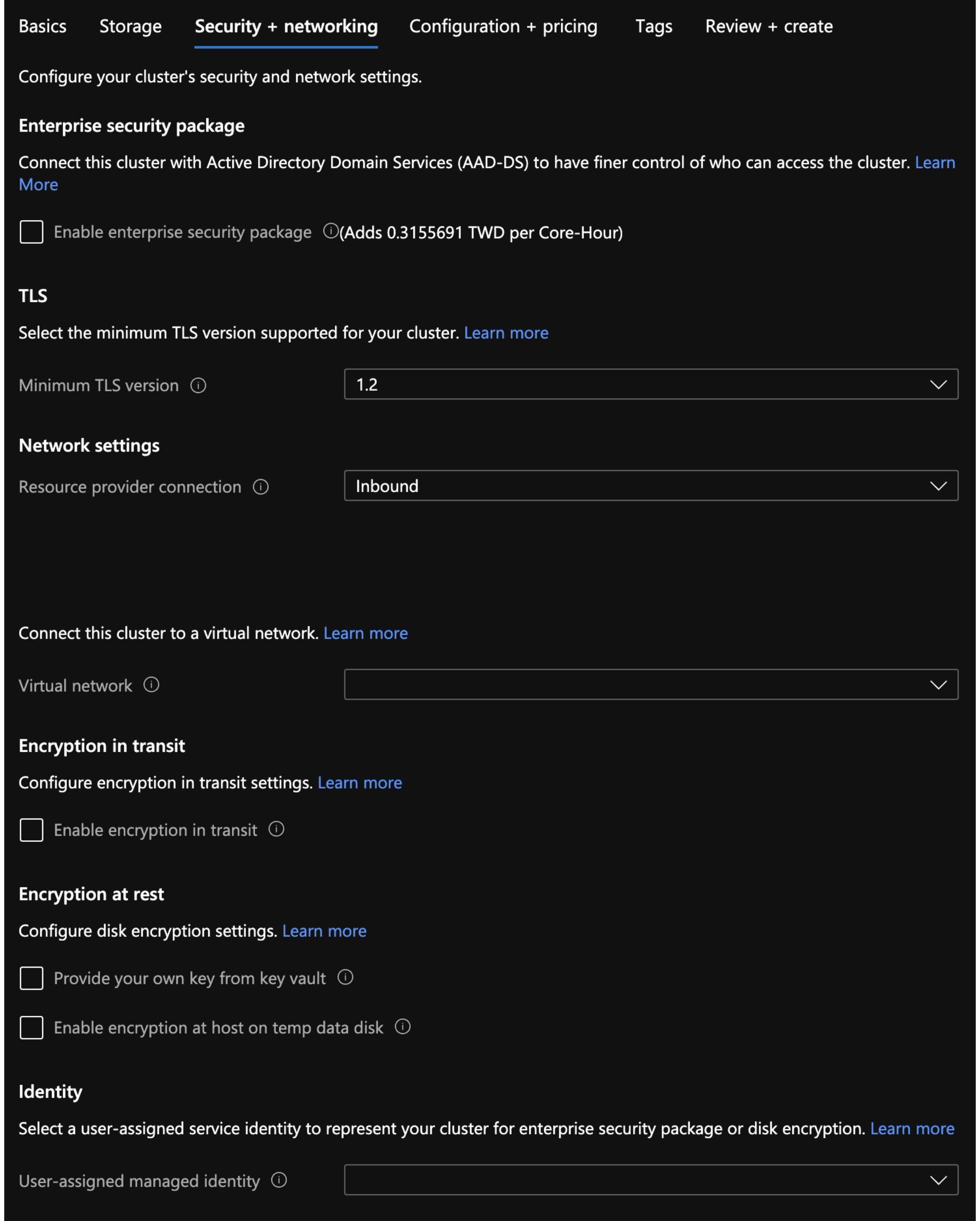
To store your Hive and Oozie metadata outside of this cluster, select a SQL database. Learn More

SQL database for Hive

SQL database for Oozie

- Choose Storage Account
- Set name for Container

Configure Security (not modified)



Configure your cluster's security and network settings.

Enterprise security package

Connect this cluster with Active Directory Domain Services (AAD-DS) to have finer control of who can access the cluster. Learn More

□ Enable enterprise security package ⓘ(Adds 0.3155691 TWD per Core-Hour)

TLS

Select the minimum TLS version supported for your cluster. Learn more

Minimum TLS version ⓘ: 1.2

Network settings

Resource provider connection ⓘ: Inbound

Connect this cluster to a virtual network. Learn more

Virtual network ⓘ:

Encryption in transit

Configure encryption in transit settings. Learn more

□ Enable encryption in transit ⓘ

Encryption at rest

Configure disk encryption settings. Learn more

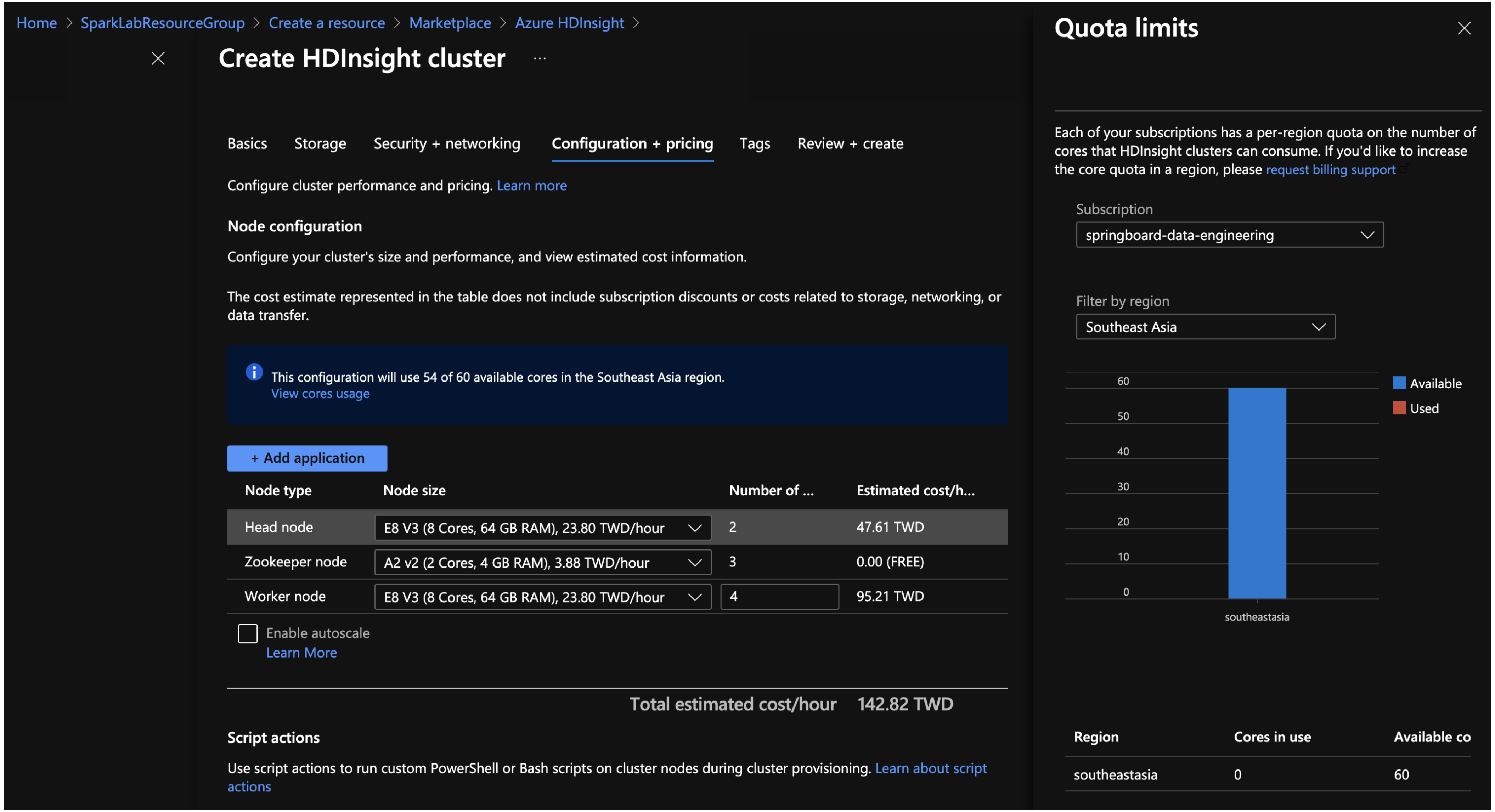
□ Provide your own key from key vault ⓘ

□ Enable encryption at host on temp data disk ⓘ

Identity

Select a user-assigned service identity to represent your cluster for enterprise security package or disk encryption. Learn more

User-assigned managed identity ⓘ:



Home > SparkLabResourceGroup > Create a resource > Marketplace > Azure HDInsight >

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 54 of 60 available cores in the Southeast Asia region.
View cores usage

+ Add application

Node type	Node size	Number of ...	Estimated cost/h...
Head node	E8 V3 (8 Cores, 64 GB RAM), 23.80 TWD/hour	2	47.61 TWD
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 3.88 TWD/hour	3	0.00 (FREE)
Worker node	E8 V3 (8 Cores, 64 GB RAM), 23.80 TWD/hour	4	95.21 TWD

□ Enable autoscale
Learn More

Total estimated cost/hour 142.82 TWD

Script actions

Use script actions to run custom PowerShell or Bash scripts on cluster nodes during cluster provisioning. Learn about script actions

Quota limits

Each of your subscriptions has a per-region quota on the number of cores that HDInsight clusters can consume. If you'd like to increase the core quota in a region, please request billing support

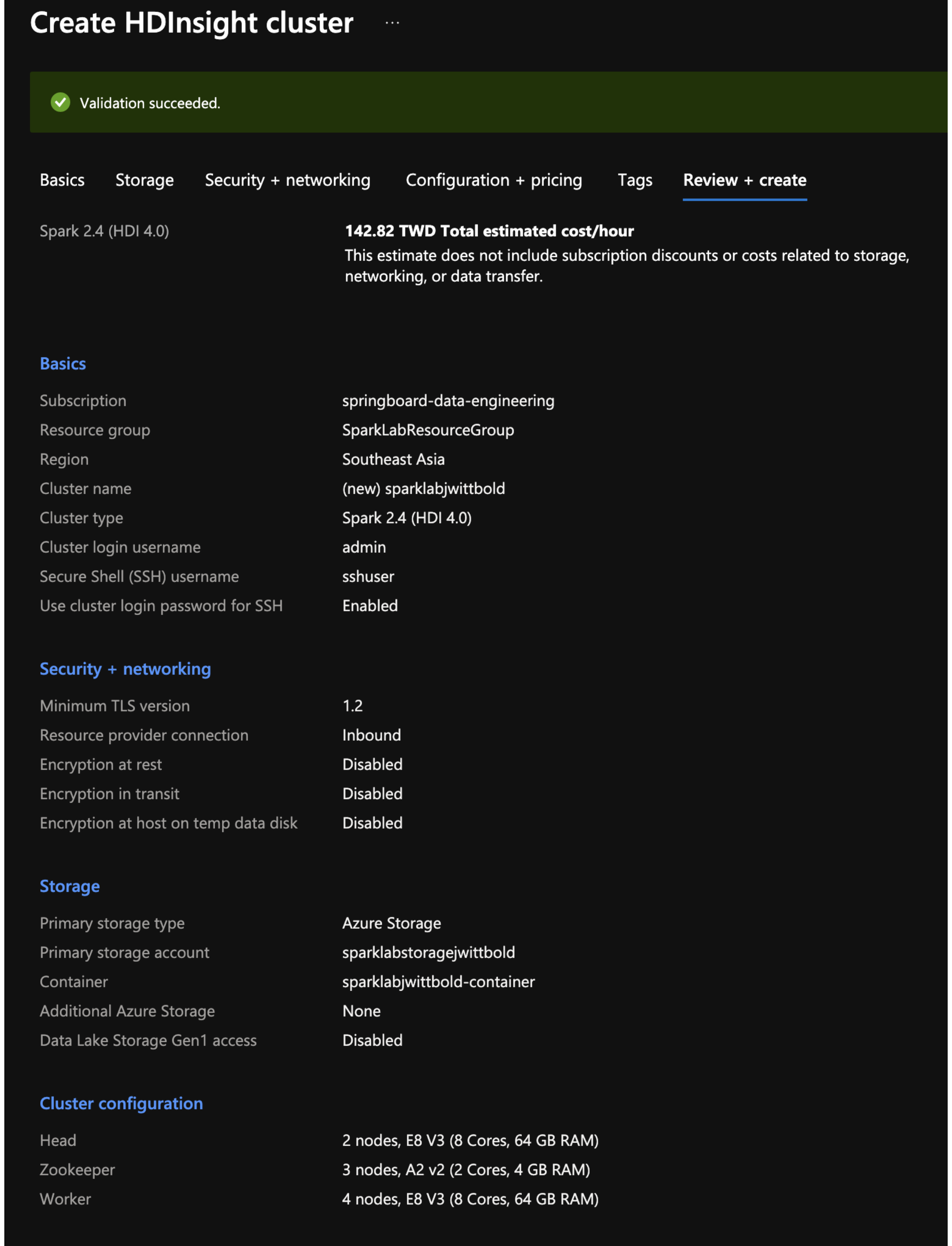
Subscription: springboard-data-engineering

Filter by region: Southeast Asia

Region	Cores in use	Available co
southeastasia	0	60

Configure Nodes

- Needed to first request Core Quota increase for subscription / region



Validation succeeded.

Spark 2.4 (HDI 4.0)

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

Basics

Subscription	springboard-data-engineering
Resource group	SparkLabResourceGroup
Region	Southeast Asia
Cluster name	(new) sparklabjwittbold
Cluster type	Spark 2.4 (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

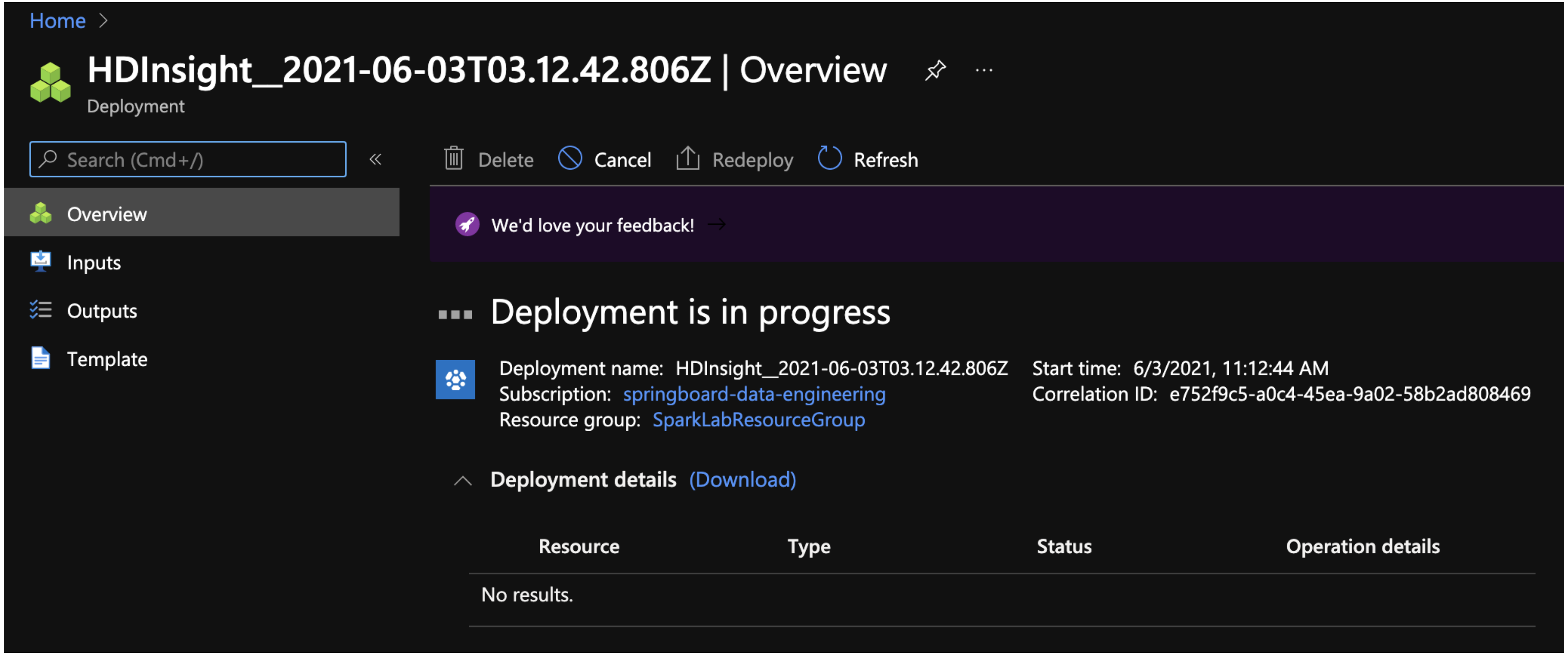
Primary storage type	Azure Storage
Primary storage account	sparklabstoragejwittbold
Container	sparklabjwittbold-container
Additional Azure Storage	None
Data Lake Storage Gen1 access	Disabled

Cluster configuration

Head	2 nodes, E8 V3 (8 Cores, 64 GB RAM)
Zookeeper	3 nodes, A2 v2 (2 Cores, 4 GB RAM)
Worker	4 nodes, E8 V3 (8 Cores, 64 GB RAM)

Full Details of HDInsight Cluster

- Validation Successful



Deployment is in progress

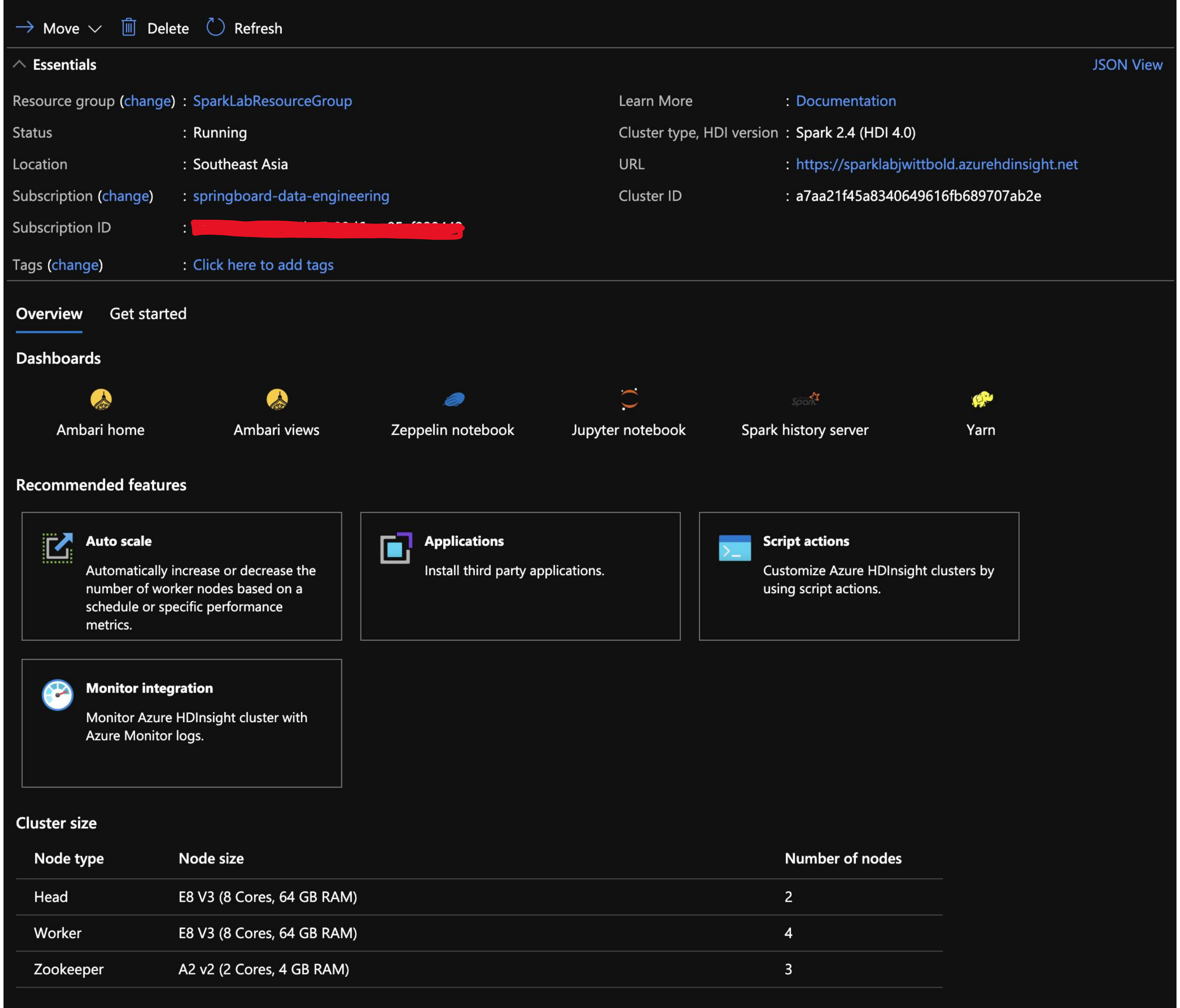
Deployment name: HDInsight__2021-06-03T03.12.42.806Z
Subscription: springboard-data-engineering
Resource group: SparkLabResourceGroup

Start time: 6/3/2021, 11:12:44 AM
Correlation ID: e752f9c5-a0c4-45ea-9a02-58b2ad808469

Deployment details (Download)

Resource	Type	Status	Operation details
No results.			

Deployment in progress...



Essentials

JSON View

Resource group (change) : SparkLabResourceGroup

Status : Running

Location : Southeast Asia

Subscription (change) : springboard-data-engineering

Subscription ID : [redacted]

Tags (change) : Click here to add tags

Learn More : Documentation

Cluster type, HDI version : Spark 2.4 (HDI 4.0)

URL : https://sparklabjwittbold.azurehdinsight.net

Cluster ID : a7aa21f45a8340649616fb689707ab2e

Overview   Get started

Dashboards

Recommended features

Auto scale

Automatically increase or decrease the number of worker nodes based on a schedule or specific performance metrics.

Applications

Install third party applications.

Script actions

Customize Azure HDInsight clusters by using script actions.

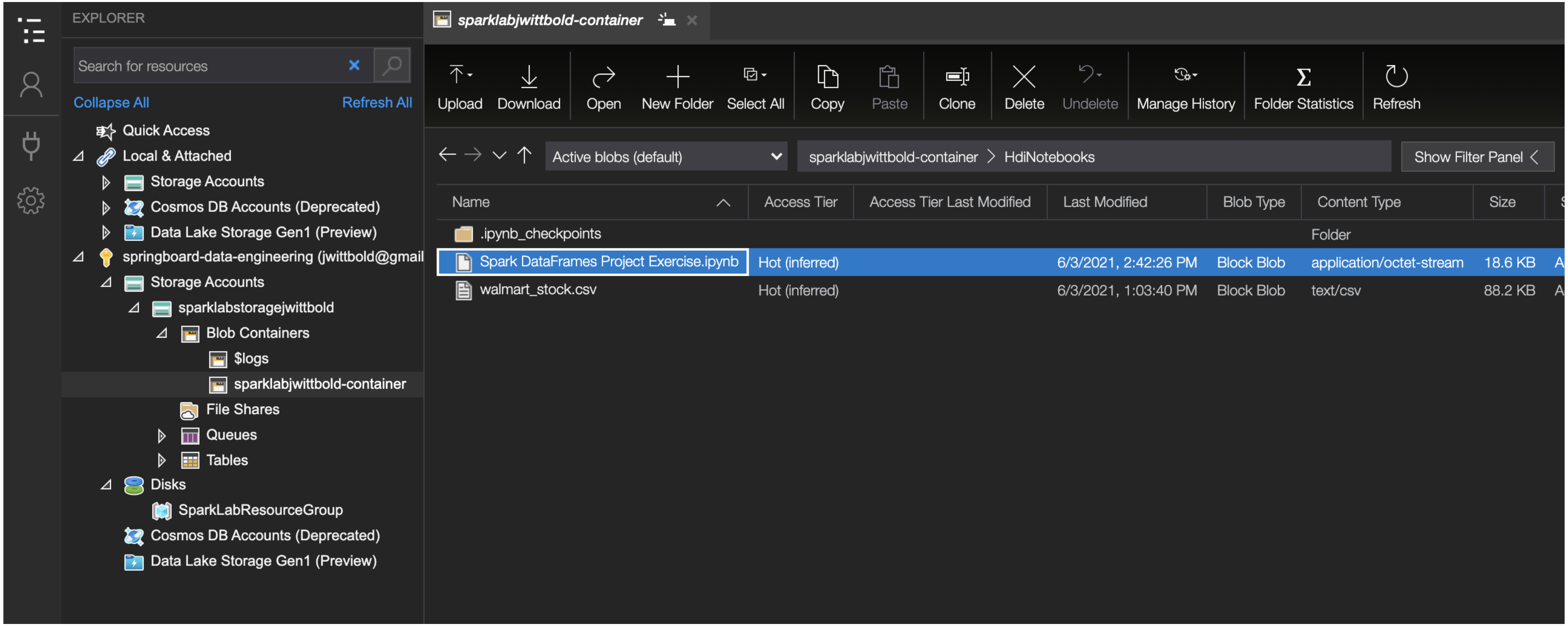
Monitor integration

Monitor Azure HDInsight cluster with Azure Monitor logs.

Cluster size

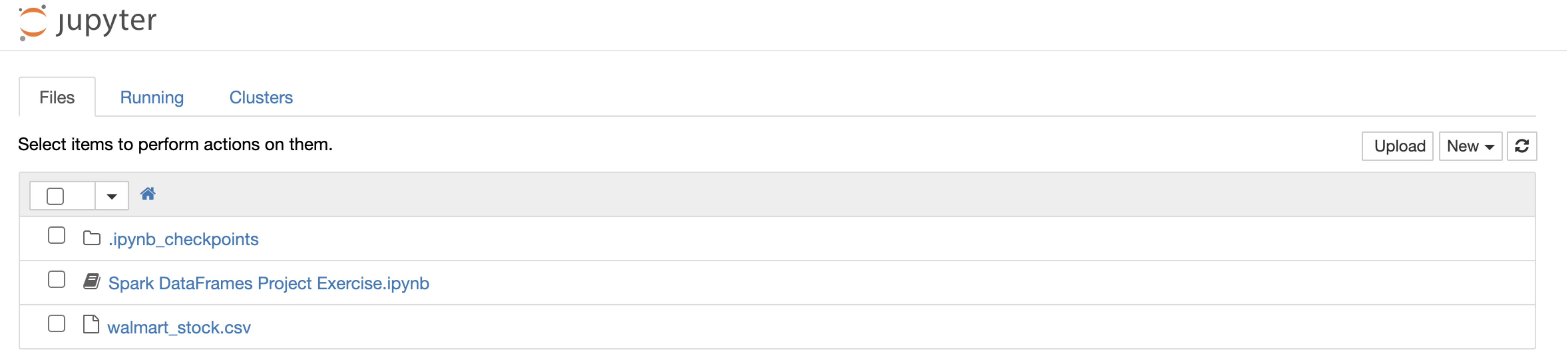
Node type	Node size	Number of nodes
Head	E8 V3 (8 Cores, 64 GB RAM)	2
Worker	E8 V3 (8 Cores, 64 GB RAM)	4
Zookeeper	A2 v2 (2 Cores, 4 GB RAM)	3

Cluster successfully created



Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size
.ipynb_checkpoints					Folder	
Spark DataFrames Project Exercise.ipynb	Hot (inferred)		6/3/2021, 2:42:26 PM	Block Blob	application/octet-stream	18.6 KB
walmart_stock.csv	Hot (inferred)		6/3/2021, 1:03:40 PM	Block Blob	text/csv	88.2 KB

Resource files uploaded to cluster via Azure Storage Explorer



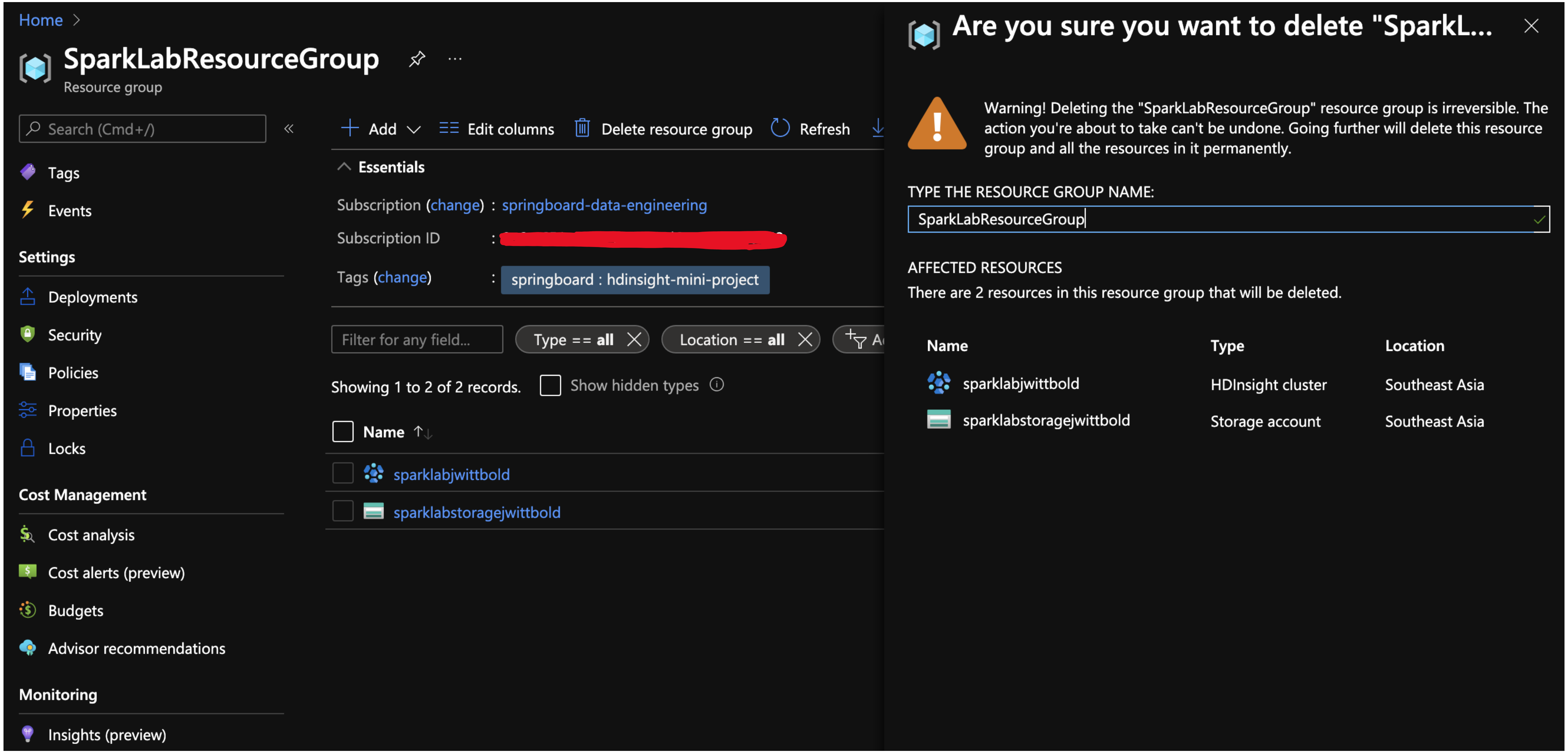
Select items to perform actions on them.

□ .ipynb_checkpoints

□ Spark DataFrames Project Exercise.ipynb

□ walmart_stock.csv

Jupyter Notebook and walmart_stock.csv data file available to work with



SparkLabResourceGroup

Resource group

Essentials

Subscription (change) : springboard-data-engineering

Subscription ID : [redacted]

Tags (change) : springboard : hdinsight-mini-project

Showing 1 to 2 of 2 records.

□ Name ↑↓
□ sparklabjwittbold
□ sparklabstoragejwittbold

Are you sure you want to delete "SparkL...

Warning! Deleting the "SparkLabResourceGroup" 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:

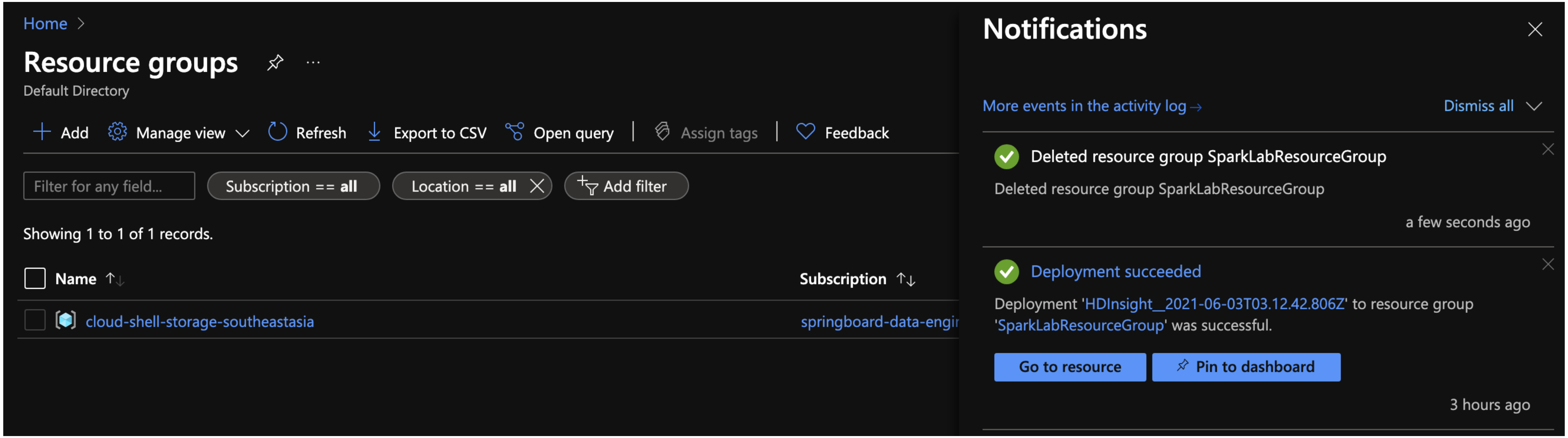
SparkLabResourceGroup ✓

AFFECTED RESOURCES

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

Name	Type	Location
sparklabjwittbold	HDInsight cluster	Southeast Asia
sparklabstoragejwittbold	Storage account	Southeast Asia

Deleting Resource Group (and all contained contents) after completing exercise



Showing 1 to 1 of 1 records.

Name	Subscription
cloud-shell-storage-southeastasia	springboard-data-engir

Deleted resource group SparkLabResourceGroup

Deleted resource group SparkLabResourceGroup

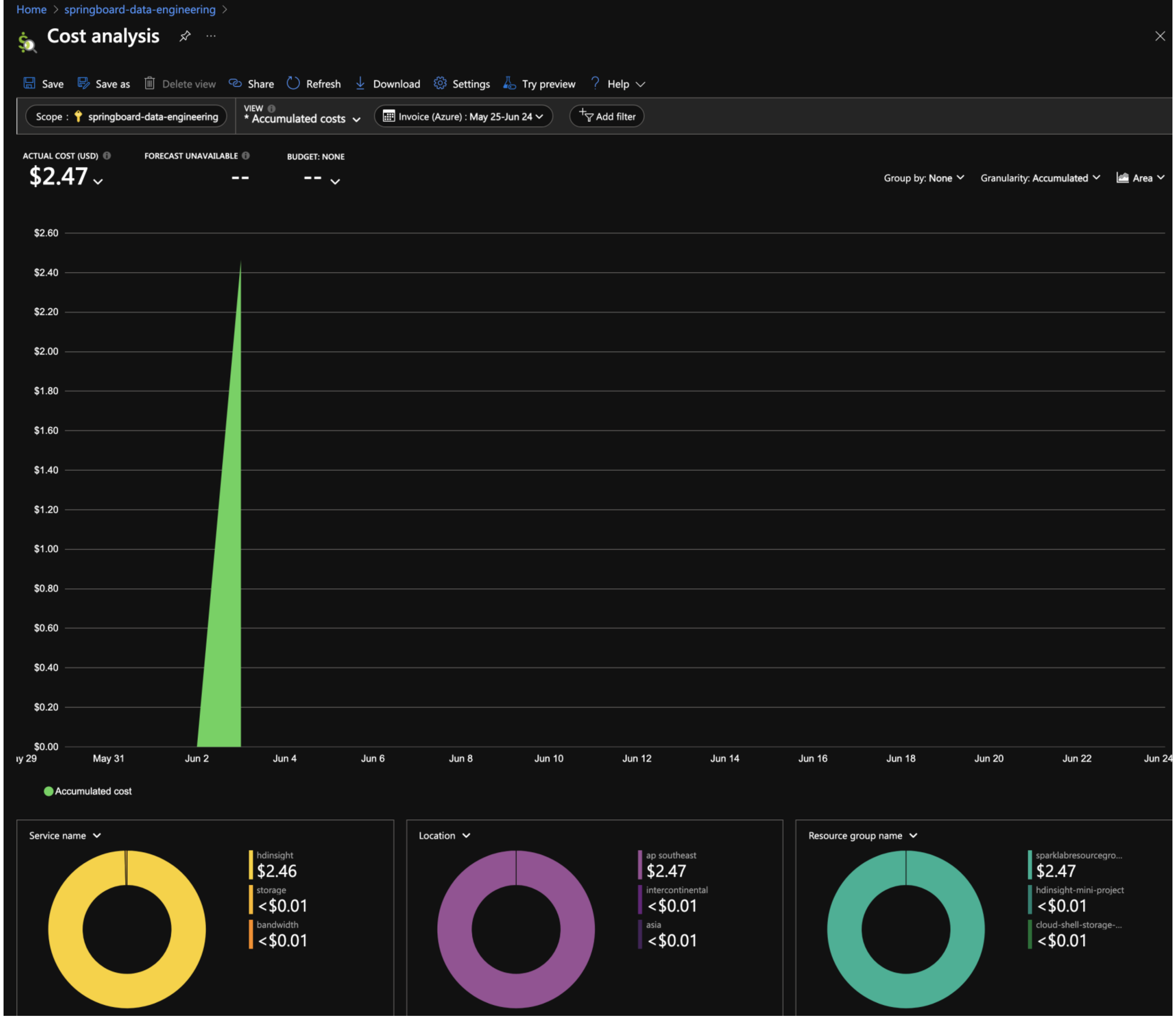
a few seconds ago

Deployment succeeded

Deployment 'HDInsight__2021-06-03T03.12.42.806Z' to resource group 'SparkLabResourceGroup' was successful.

3 hours ago

Successfully deleted



Home > springboard-data-engineering >

Cost analysis

Save | Save as | Delete view | Share | Refresh | Download | Settings | Try preview | Help

Scope : springboard-data-engineering

VIEW: * Accumulated costs

Invoice (Azure) : May 25-Jun 24

Add filter

ACTUAL COST (USD) $2.47

FORECAST UNAVAILABLE --

BUDGET: NONE --

Group by: None

Granularity: Accumulated

Area

Service name

hdinsight $2.46
storage <$0.01
bandwidth <$0.01

Location

ap southeast $2.47
intercontinental <$0.01
asia <$0.01

Resource group name

sparklabresourcegro... $2.47
hdinsight-mini-project <$0.01
cloud-shell-storage-... <$0.01

Cost Analysis for ~ 3hr HDInsight Cluster

= $2.47 USD