

Setting up your HDI 3.6 (Spark 2.1) Cluster

Configuring Basic Settings

Quick create

Custom (size, settings, apps)

1 Basics
Configure basic settings

2 Storage
Set storage settings

3 Summary
Confirm configurations

i

This cluster may take up to 20 minutes to create.

* Cluster name

bigdl

.azurehdinsight.net

* Subscription

ADLS Performance Test

* Cluster type ⓘ

Spark 2.0 on Linux (HDI 3.5)

* Cluster login username ⓘ

admin

* Cluster login password

.....

Secure Shell (SSH) username ⓘ

sshuser

☐ Use same password as cluster login ⓘ

SSH authentication type

PASSWORD PUBLIC KEY

SSH authentication type

PASSWORD PUBLIC KEY

* SSH public key

w/Aes584h+7XkQR7kGHKhM9+DKcsYZUjV
7maE/xTyzXooEZ8fD/NCmywinBmzYE3c+IJe
f/+Fd3zHDHsIfkG+6cJkyvqDUcl/u99uupf
yulvy0dNIfROBasglZk9fzGcUvGFFFc++eFGn
GHKNNQKvtwme8lNCvi2E2ch69sJM54C0D/
IAZIKzKHSWferruT5VK0IAFJ3dSLxysklovOyq
V0oBNbHkJt2XaNGbb9eDH3f

Select a file

* Resource group ⓘ

☐ Create new ☒ Use existing

doctorwho

* Location

Central US

i

Click here to view cores usage.

Configuring Basic Settings

Quick create

Custom (size, settings, apps)

1

Basics

Configure basic settings

✓

2

Storage

Set storage settings

>

3

Summary

Confirm configurations

>

i

This cluster may take up to 20 minutes to create.

i

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

Storage Account Settings

* Primary storage type

Azure Storage

Data Lake Store

* Selection method ⓘ

My subscriptions

Access key

* Select a Storage account

hdiadltest01cus

>

Create new

* Default container ⓘ

bigdl-2017-06-01t04-37-54-225z

Additional storage accounts

Optional

>

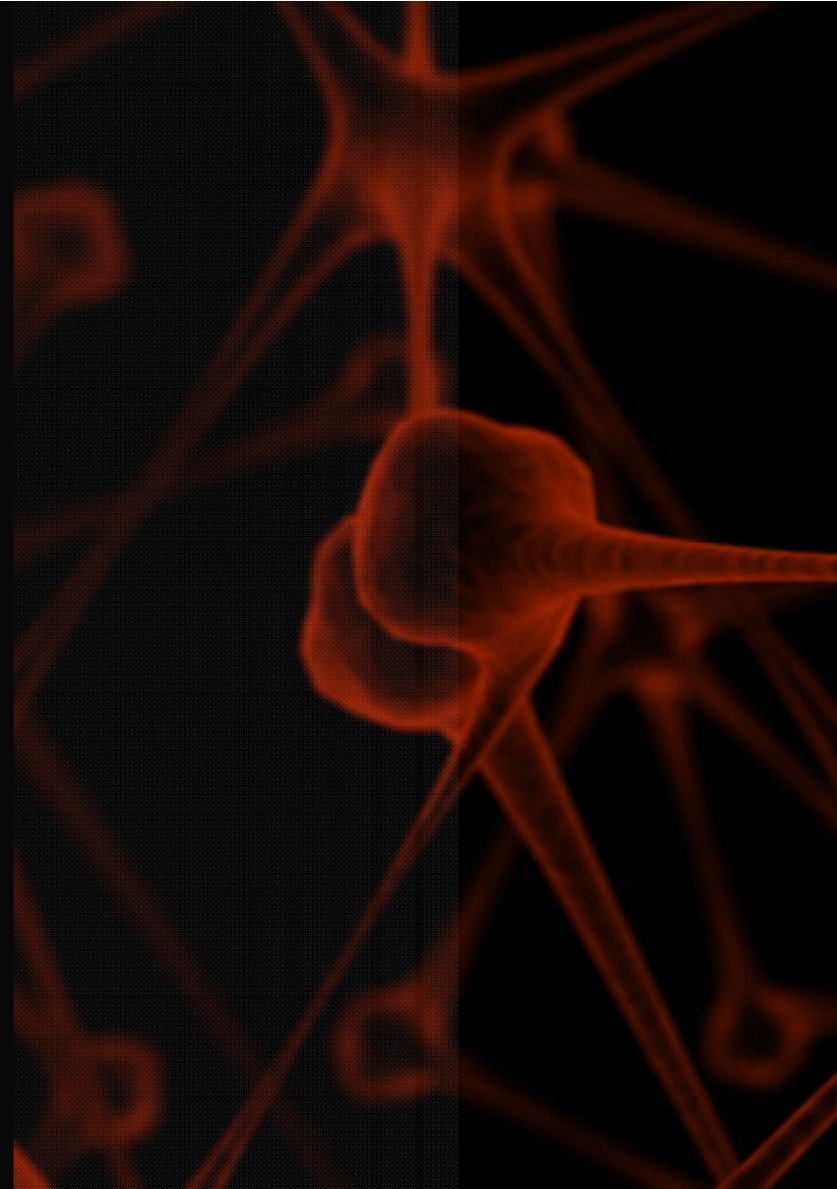
Data Lake Store access ⓘ

Optional

>

Remember this!

Tutorial Setup



Tutorial Prerequisites

- Setup HDI 3.6 (Spark 2.1) Cluster
- Copy JAR, Data, and notebooks from <http://aka.ms/BigDLTutorial> to a local folder
 - Copy the JAR to default storage
 - Copy the data to a new folder (e.g. /tmp)
 - Upload the notebook

Copy the BigDL JAR to your default storage account

Upload the BigDL JAR from Github (aka.ms/BigDLTutorial > jars folder) to your HDI cluster's default storage root (/) folder. If you're using HDI 3.6, copy the 2.1 version; if you're using HDI 3.5, copy the 2.0 version.

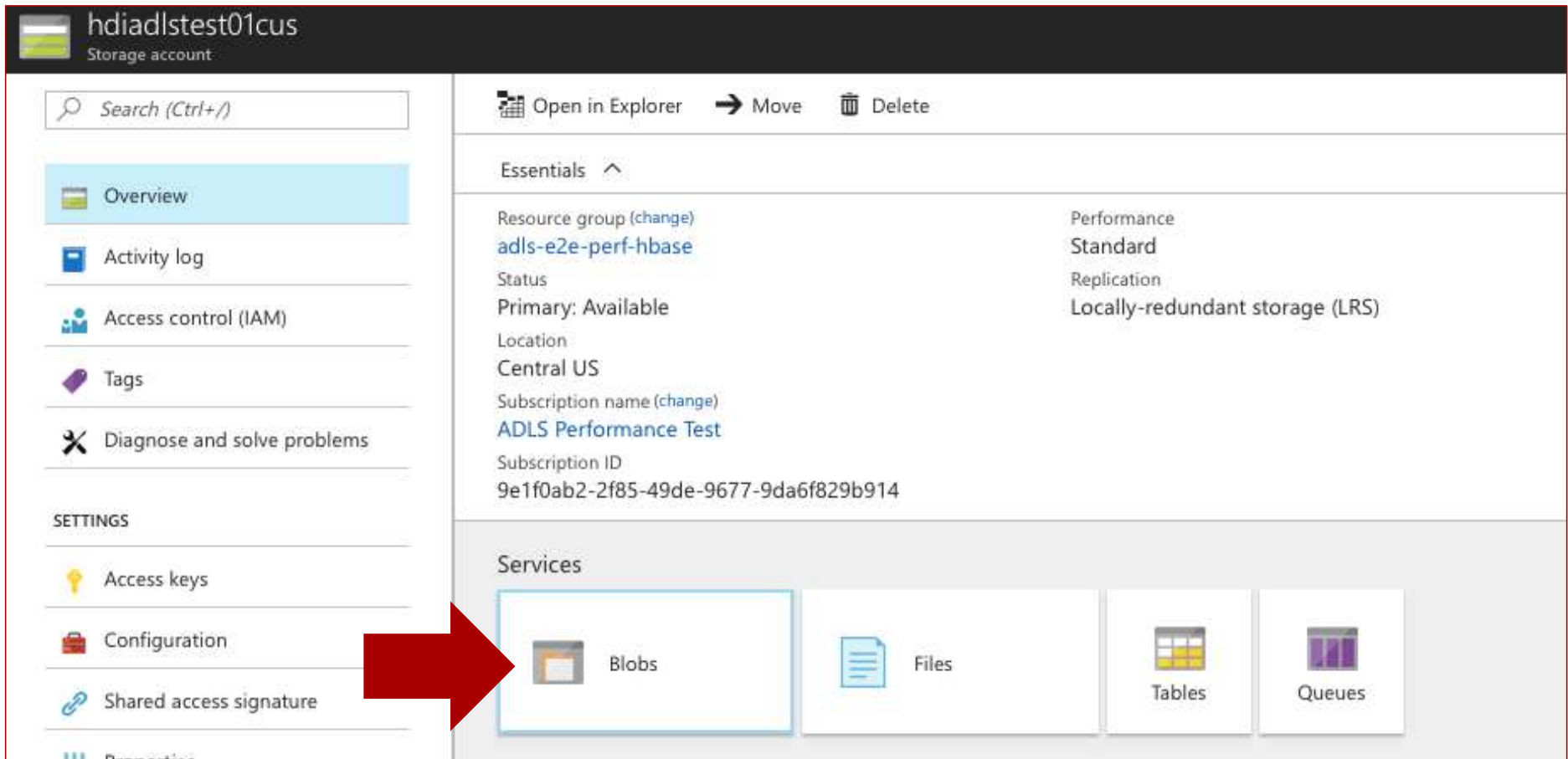
Go to your default storage account

The screenshot shows the 'bigdl - Storage accounts' window. The left sidebar contains a search bar and several menu items: 'External metastores', 'Script actions', 'Monitoring', 'PROPERTIES' (with sub-items 'Properties', 'Storage accounts', and 'Data Lake Store access'), and 'SUPPORT + TROUBLESHOOTING' (with sub-items 'Resource health' and 'New support request'). The 'Storage accounts' item is highlighted. The main area displays a table with the following data:

STORAGE	CONTAINER / DIRECTORY	DEFAULT
hdiadltest01cus	bigdl-2017-06-01t04-37-54-225z	<input checked="" type="checkbox"/>

Red arrows with numbers 1 and 2 indicate the steps to reach the default storage account. Arrow 1 points to the 'Storage accounts' option in the sidebar, and arrow 2 points to the 'hdiadltest01cus' storage account in the table.

Go to your default storage account (cont'd)



The screenshot displays the Azure portal interface for a storage account named **hdiadltest01cus**. The left-hand navigation pane includes sections for **Overview**, **Activity log**, **Access control (IAM)**, **Tags**, **Diagnose and solve problems**, and **SETTINGS** (which contains **Access keys**, **Configuration**, and **Shared access signature**). The main content area features a search bar, action buttons (**Open in Explorer**, **Move**, **Delete**), and an **Essentials** section. The Essentials section lists account details: Resource group ([change](#)) **adls-e2e-perf-hbase**, Status **Primary: Available**, Location **Central US**, Subscription name ([change](#)) **ADLS Performance Test**, and Subscription ID **9e1f0ab2-2f85-49de-9677-9da6f829b914**. It also shows Performance **Standard** and Replication **Locally-redundant storage (LRS)**. Below this is the **Services** section, which contains four tiles: **Blobs**, **Files**, **Tables**, and **Queues**. A large red arrow points from the **Configuration** link in the left sidebar to the **Blobs** tile.

hdiadltest01cus
Storage account

Search (Ctrl+/)

Open in Explorer → Move Delete

Essentials ^

Resource group ([change](#))
adls-e2e-perf-hbase

Status
Primary: Available

Location
Central US

Subscription name ([change](#))
ADLS Performance Test

Subscription ID
9e1f0ab2-2f85-49de-9677-9da6f829b914

Performance
Standard

Replication
Locally-redundant storage (LRS)

Services

Blobs **Files** **Tables** **Queues**

Go to your default storage account (cont'd)

Blob service

hdiadltest01cus

+ Container

Refresh

Essentials ^

Storage account

hdiadltest01cus

Status

Primary: Available

Location

Central US

Subscription (change)

ADLS Performance Test

Subscription ID

9e1f0ab2-2f85-49de-9677-9da6f829b914

Blob service endpoint

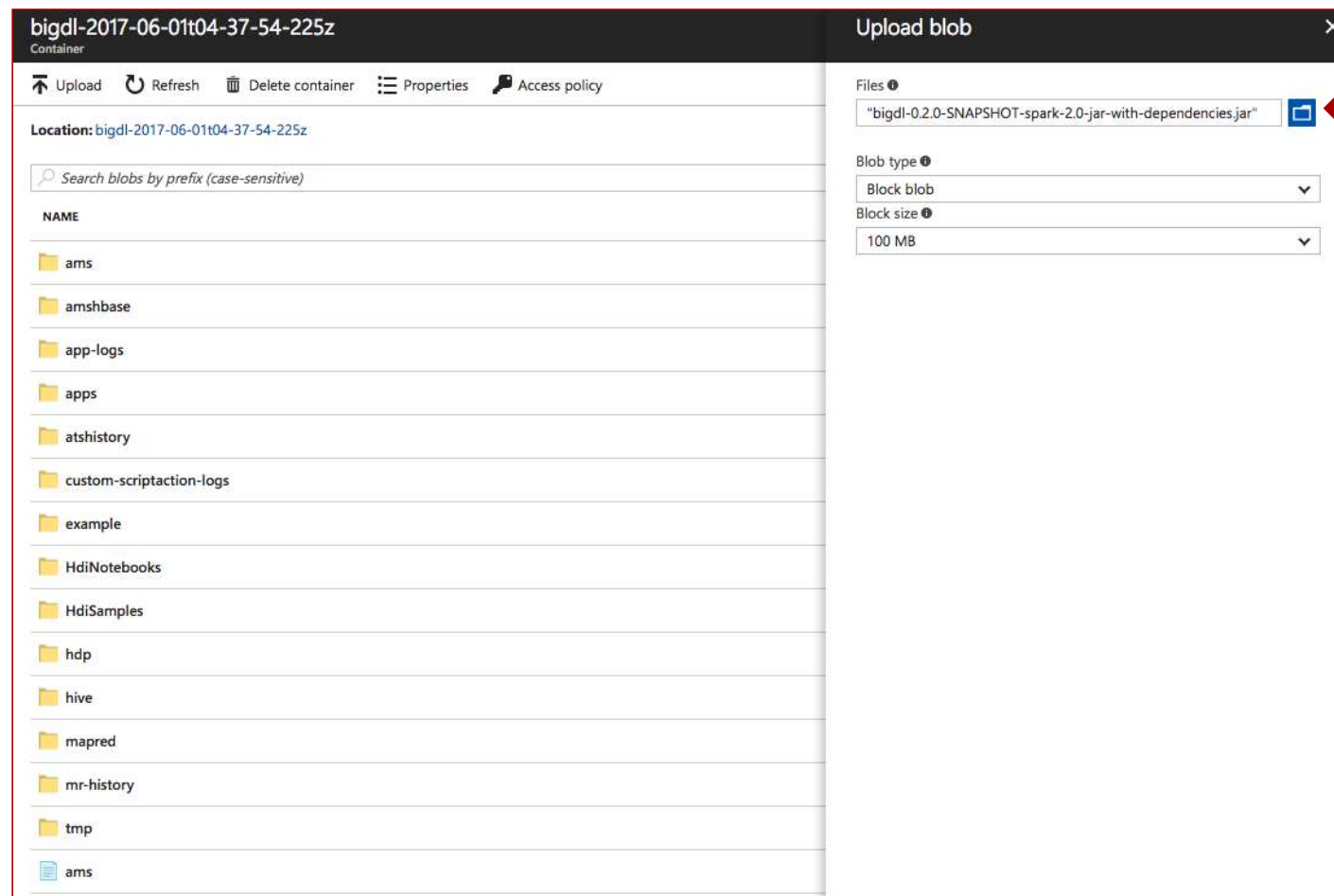
https://hdiadltest01cus.blob.core.windows.net/

Search containers by prefix

NAME	LAST MODIFIED	ACCESS TYPE	LEASE STATE
bigdl-2017-06-01t04-37-54-225z	5/31/2017, 10:12:21 PM	Private	Available ...
cxcusmhteffa02e57db14929aec9b5d748ce18f1	4/7/2017, 9:39:32 AM	Private	Available ...



Upload the BigDL JAR to your storage account



The screenshot displays the Azure Storage portal interface. On the left, a container named 'bigdl-2017-06-01t04-37-54-225z' is shown with a list of blobs. On the right, the 'Upload blob' dialog box is open. The 'Files' field contains the filename 'bigdl-0.2.0-SNAPSHOT-spark-2.0-jar-with-dependencies.jar'. A red arrow points to this field. The 'Blob type' is set to 'Block blob' and the 'Block size' is set to '100 MB'.

bigdl-2017-06-01t04-37-54-225z
Container


Upload Refresh Delete container Properties Access policy

Location: bigdl-2017-06-01t04-37-54-225z

Search blobs by prefix (case-sensitive)

NAME
ams
amshbase
app-logs
apps
atshistory
custom-scriptaction-logs
example
HdiNotebooks
HdiSamples
hdp
hive
mapred
mr-history
tmp
ams

Upload blob

Files ⓘ
"bigdl-0.2.0-SNAPSHOT-spark-2.0-jar-with-dependencies.jar" 

Blob type ⓘ
Block blob ▼

Block size ⓘ
100 MB ▼

Upload MNIST data

Upload the MNIST data from Github (aka.ms/BigDLTutorial > data folder) to your HDI cluster's default storage /tmp folder

Upload MNIST data

Blob service
hdiadltest01cus

+

 Container

↺

 Refresh

Essentials ▾

🔍 Search containers by prefix

NAME

bigdl-2017-06-01t04-37-54-225z ...

cxcusmhteffa02e57db14929aec9b5d... ...

cxcusmhteffa02e57db14929aec9b5d... ...

felixtest01-2017-05-28t01-43-18-595z ...

testfelix11-2017-05-12t22-58-32-966z ...

testfelixfoo-2017-05-26t18-09-09-95... ...

bigdl-2017-06-01t04-37-54-225z
Container

📁 Upload

↺

 Refresh

🗑️

 Delete container

⋮

 Properties

🔑

 Access policy

📄 apps	5/31/2017, 10:02:01 PM	Block blob	0 B
📄 atshistory	5/31/2017, 10:02:01 PM	Block blob	0 B
📄 bigdl-0.2.0-SNAPSHOT-spark-2.0-jar-with-dependencies.jar	6/1/2017, 8:08:35 PM	Block blob	54.8 MiB
📄 custom-scriptaction-logs	5/31/2017, 10:15:54 PM	Block blob	0 B
📄 example	5/31/2017, 10:15:00 PM	Block blob	0 B
📄 hbase	5/31/2017, 10:02:01 PM	Block blob	0 B
📄 Hd>Notebooks	5/31/2017, 10:12:46 PM	Block blob	0 B
📄 Hd:Samples	5/31/2017, 10:15:18 PM	Block blob	0 B
📄 hdp	5/31/2017, 10:02:01 PM	Block blob	0 B
📄 hive	5/31/2017, 10:02:01 PM	Block blob	0 B
📄 mapred	5/31/2017, 10:02:02 PM	Block blob	0 B
📄 mr-history	5/31/2017, 10:02:02 PM	Block blob	0 B
📄 tmp	5/31/2017, 10:02:02 PM	Block blob	0 B
📁 tmp			-
📁 user			-

Upload MNIST data

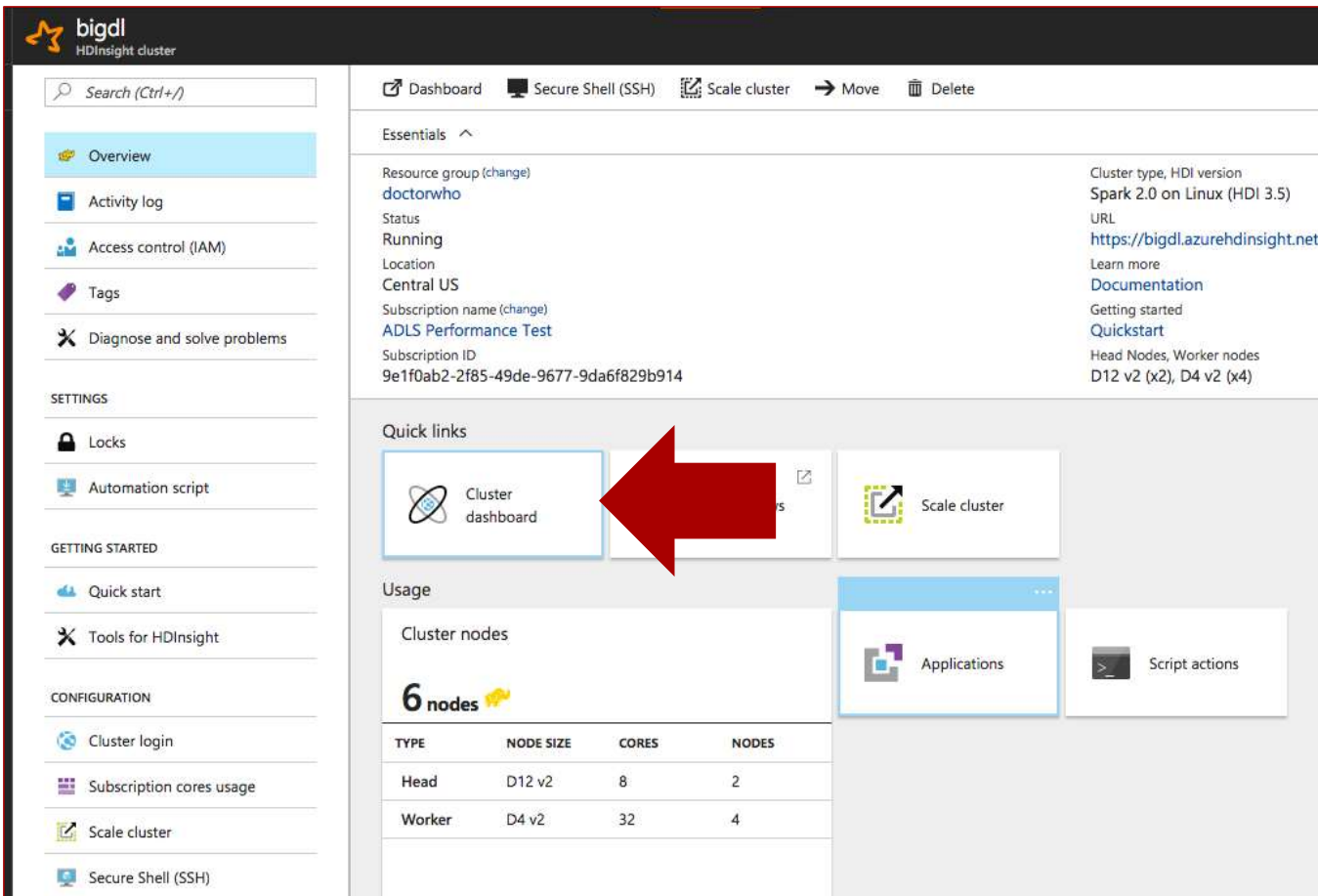
The screenshot displays the Azure Blob Storage management interface. It is divided into three main sections: a left sidebar for navigation, a central pane for the selected container, and a right pane for uploading new blobs.

- Left Sidebar:** Shows the 'Blob service' for 'hdiadltest01cus'. Under 'Essentials', a search bar is present. A list of containers is shown, with 'bigdl-2017-06-01t04-37-54-225z' selected and highlighted in blue.
- Central Pane:** Displays the contents of the 'tmp' folder. It includes an 'Upload' button and a 'Refresh' button. Below these, a list of blobs is shown, including a folder icon for '[..]' and a file icon for 'entity-file-history'. A red arrow points to the 'Upload' button.
- Right Pane:** Titled 'Upload blob', it contains a 'Files' section with a text input field containing the path '"t10k-images-idx3-ubyte" "t10k-labels-idx1-ubyte" "train-ima...'. Below this, there are dropdown menus for 'Blob type' (set to 'Block blob') and 'Block size' (set to '100 MB'). A red arrow points to the top right corner of this pane.

Upload the notebook

Upload the MNIST notebook from Github (aka.ms/BigDLTutorial) to your Jupyter notebook service Scala folder.

Open Jupyter notebook



bigdl
HDInsight cluster

Search (Ctrl+/)

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

SETTINGS

- Locks
- Automation script

GETTING STARTED

- Quick start
- Tools for HDInsight

CONFIGURATION

- Cluster login
- Subscription cores usage
- Scale cluster
- Secure Shell (SSH)

Dashboard | Secure Shell (SSH) | Scale cluster | Move | Delete

Essentials

Resource group (change): doctorwho
Status: Running
Location: Central US
Subscription name (change): ADLS Performance Test
Subscription ID: 9e1f0ab2-2f85-49de-9677-9da6f829b914

Cluster type, HDI version: Spark 2.0 on Linux (HDI 3.5)
URL: <https://bigdl.azurehdinsight.net>
Learn more: [Documentation](#)
Getting started: [Quickstart](#)
Head Nodes, Worker nodes: D12 v2 (x2), D4 v2 (x4)

Quick links

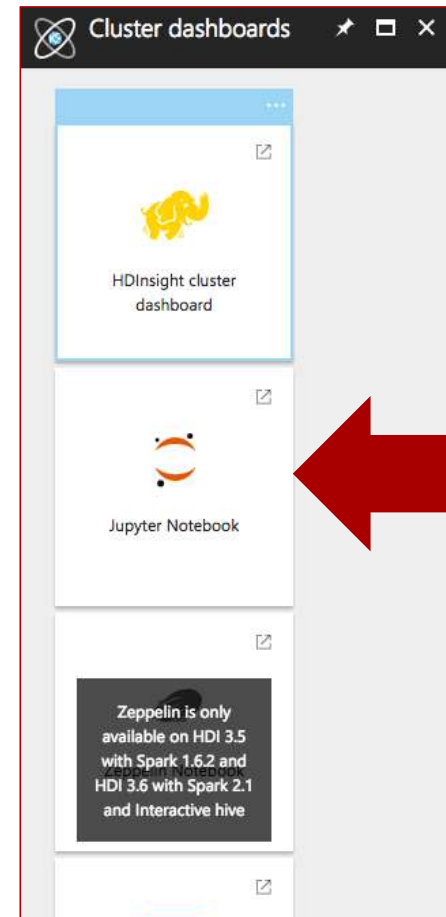
- Cluster dashboard** (highlighted with a red arrow)
- Scale cluster

Usage

Cluster nodes: **6 nodes**

TYPE	NODE SIZE	CORES	NODES
Head	D12 v2	8	2
Worker	D4 v2	32	4

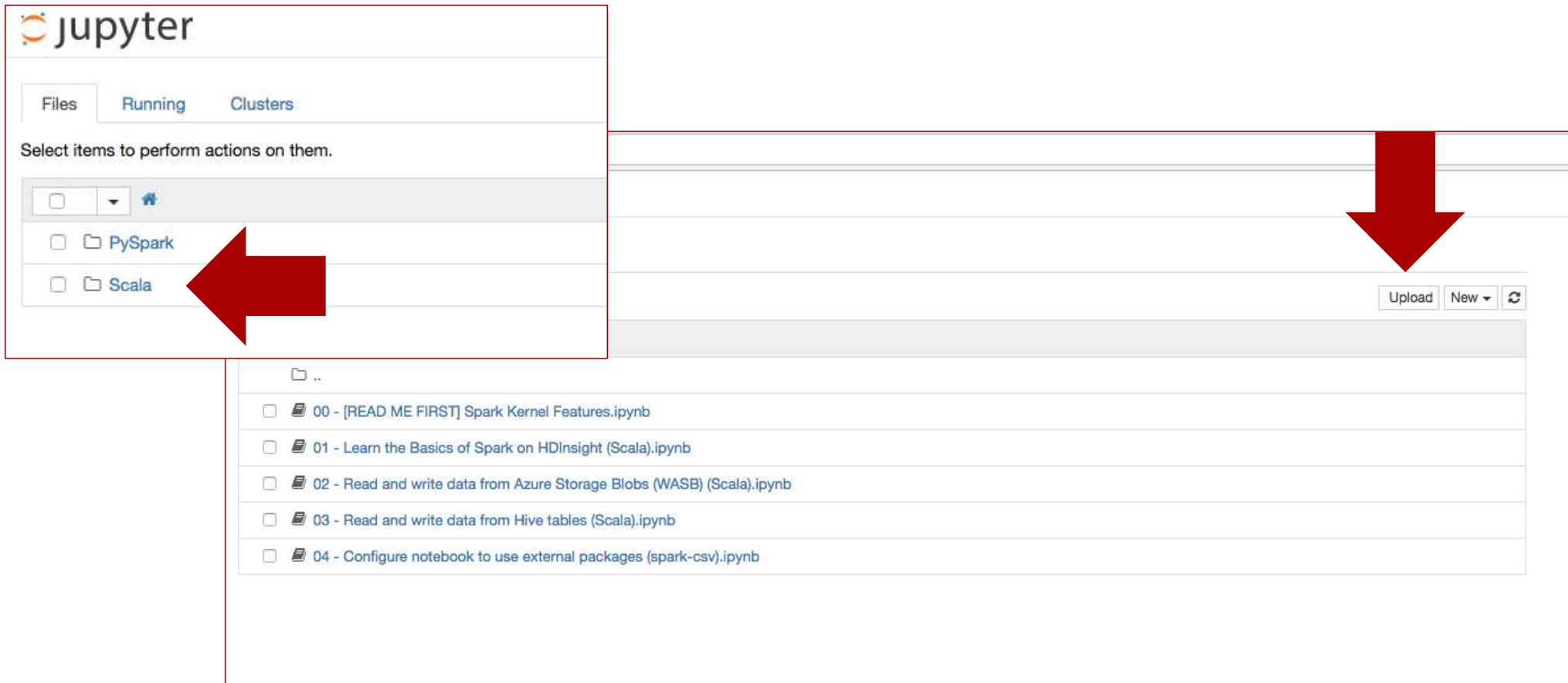
Applications | **Script actions**



Cluster dashboards

- HDInsight cluster dashboard
- Jupyter Notebook** (highlighted with a red arrow)
- Zeppelin is only available on HDI 3.5 with Spark 1.6.2 and HDI 3.6 with Spark 2.1 and Interactive hive

Upload Scala notebook



The image shows the Jupyter web interface. On the left, the 'Files' tab is active, displaying a file browser. A red arrow points to the 'Scala' folder. On the right, a large red arrow points down to the 'Upload' button. Below the file browser, a list of notebooks is visible.

Jupyter Files Tab:

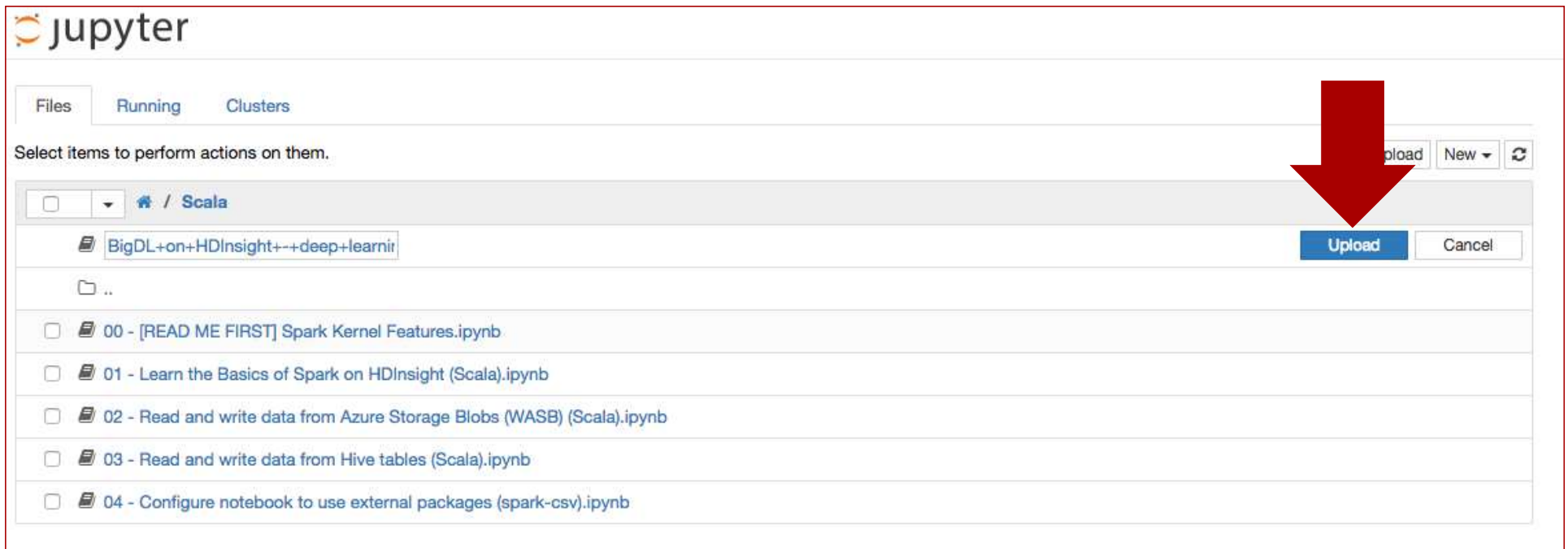
- Select items to perform actions on them.
- ☐ ..
- ☐ PySpark
- ☐ Scala

Notebook List:

- ☐ 00 - [READ ME FIRST] Spark Kernel Features.ipynb
- ☐ 01 - Learn the Basics of Spark on HDInsight (Scala).ipynb
- ☐ 02 - Read and write data from Azure Storage Blobs (WASB) (Scala).ipynb
- ☐ 03 - Read and write data from Hive tables (Scala).ipynb
- ☐ 04 - Configure notebook to use external packages (spark-csv).ipynb

Buttons: Upload, New,

Upload Scala notebook (continued)



The image shows the Jupyter web interface's 'Files' tab. At the top, there are tabs for 'Files', 'Running', and 'Clusters'. Below them, a message says 'Select items to perform actions on them.' The main area displays a file browser for the 'Scala' directory. It contains a search bar, a breadcrumb path, and a list of files. A red arrow points to the 'Upload' button in the top right corner of the file list.

jupyter

Files Running Clusters

Select items to perform actions on them.

Upload New ↕

☐ / Scala

☐ BigDL-on-HDInsight--deep-learnir

☐ ..

☐ 00 - [READ ME FIRST] Spark Kernel Features.ipynb

☐ 01 - Learn the Basics of Spark on HDInsight (Scala).ipynb

☐ 02 - Read and write data from Azure Storage Blobs (WASB) (Scala).ipynb

☐ 03 - Read and write data from Hive tables (Scala).ipynb

☐ 04 - Configure notebook to use external packages (spark-csv).ipynb

Upload Cancel