

Práctica Apache Spark

Cindy López

Práctica Apache Spark

databricks



- Compañía fundada por los creadores de Apache Spark.
- Proporciona una plataforma web para trabajar con Spark, aportando administración de los clústeres de Spark y notebooks al estilo IPython.

Práctica Apache Spark

Probando DATABRICKS PLATFORM – FREE TRIAL



14 días de prueba
(Excluidos cargos de AWS)

- Número de clústeres ilimitados escalables.
- Programador de tareas para ejecutar jobs.
- Notebooks interactivos con colaboración, dashboard y REST APIs.
- Seguridad avanzada basada en un modelo de autorización por roles y logs de auditoría.
- Desplegado en AWS.
- <https://accounts.cloud.databricks.com/registration.html#signup>

Práctica Apache Spark

Probando DATABRICKS COMMUNITY EDITION





GRATUITO.

- Un solo clúster limitado a 6GB de RAM sin workers nodes.
- Notebook básico sin posibilidad de colaboración.
- Limitado a 3 usuarios.
- Entorno público para publicar tu trabajo.
- <https://accounts.cloud.databricks.com/registration.html#signup/community>

Práctica Apache Spark

Alta en DATABRICKS COMMUNITY EDITION

Sign Up for Databricks Community Edition

First Name *	Last Name *
<input type="text" value="Santiago"/>	<input type="text" value="Monrobé Gutiérrez"/>
Company Name *	Work Email *
<input type="text" value="Buceando"/>	<input type="text" value="santiago@buceandoenlamemoria.com"/>
Password *	Confirm Password *
<input type="password" value="*****"/>	<input type="password" value="*****"/>
Phone Number	What is your intended use case? *
<input type="text" value="55555555555"/>	<input type="text" value="Personal - Learning Spark"/>
How would you describe your role? *	
<input type="text" value="Student"/>	
<div><div> No soy un robot</div><div> reCAPTCHA Privacidad - Condiciones</div></div>	
<div>Sign Up</div>	

Práctica Apache Spark

Alta en DATABRICKS COMMUNITY EDITION

Terms of Service

Services and these Terms create no third party beneficiary rights.

Termination, Modification, Waiver & Assignment. Either of us may terminate your use of the Services at any time and for any reason; however, obligations of these Terms that by their nature should survive termination shall so survive. In addition, we may revise these Terms from time to time, and will always post the most current version on our website. If we elect to terminate your account, or if a revision of these Terms meaningfully reduces your rights, we will make a reasonable attempt to notify you (by, for example, sending a message to the email address associated with your account or posting for a reasonable time period a message to the login page of the Services). By continuing to use or access the Services after the revisions come into effect, you agree to be bound by the revised Terms. Databricks' failure to enforce a provision of these Terms is not a waiver of its right to do so later. If a provision is found unenforceable, the remaining provisions of the Terms will remain in full effect and an enforceable term will be substituted reflecting our intent as closely as possible. You may not assign any of your rights under these Terms, and any such attempt will be void. Databricks may assign its rights to any of its affiliates or to any successor in interest.

BY CLICKING "AGREE" BELOW YOU ARE AGREEING THAT YOU HAVE CAREFULLY READ ALL OF THE ABOVE PROVISIONS, INCLUDING THE RISKS AND RESTRICTIONS RELATING TO YOUR DATA AND TO YOUR USE OF THE SERVICES, AND YOU ARE AGREEING TO BE BOUND BY ALL OF THE ABOVE PROVISIONS. IF YOU DO NOT AGREE WITH THESE TERMS OR DO NOT UNDERSTAND THEM, DO NOT CLICK AGREE.

[Download a print version](#)

Cancel

Agree

Práctica Apache Spark

Alta en DATABRICKS COMMUNITY EDITION

The screenshot shows the Databricks Community Edition dashboard. On the left is a dark sidebar with navigation icons and labels: Databricks logo, Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main content area has a white background. At the top right, there's an 'Upgrade' button, a help icon, and a user profile icon. The main heading is 'Welcome to databricks™'. Below this are three large cards: 'Explore the Quickstart Tutorial' (with a document icon and a lightbulb), 'Import & Explore Data' (with a dashed box icon and a cloud upload icon), and 'Create a Blank Notebook' (with a document icon and a plus sign). Each card has a brief description. Below these cards are three sections: 'Common Tasks' (listing New Notebook, Create Table, New Cluster, New Job, New MLflow Experiment with a 'New' badge, Import Library, and Read Documentation), 'Recents' (stating 'Recent files appear here as you work.'), and 'What's new in v3.9' (with a link to 'View latest release notes').

databricks

Upgrade ?

Welcome to databricks™

Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

Common Tasks

- New Notebook
- Create Table
- New Cluster
- New Job
- New MLflow Experiment New
- Import Library
- Read Documentation

Recents

Recent files appear here as you work.

What's new in v3.9

[View latest release notes](#)

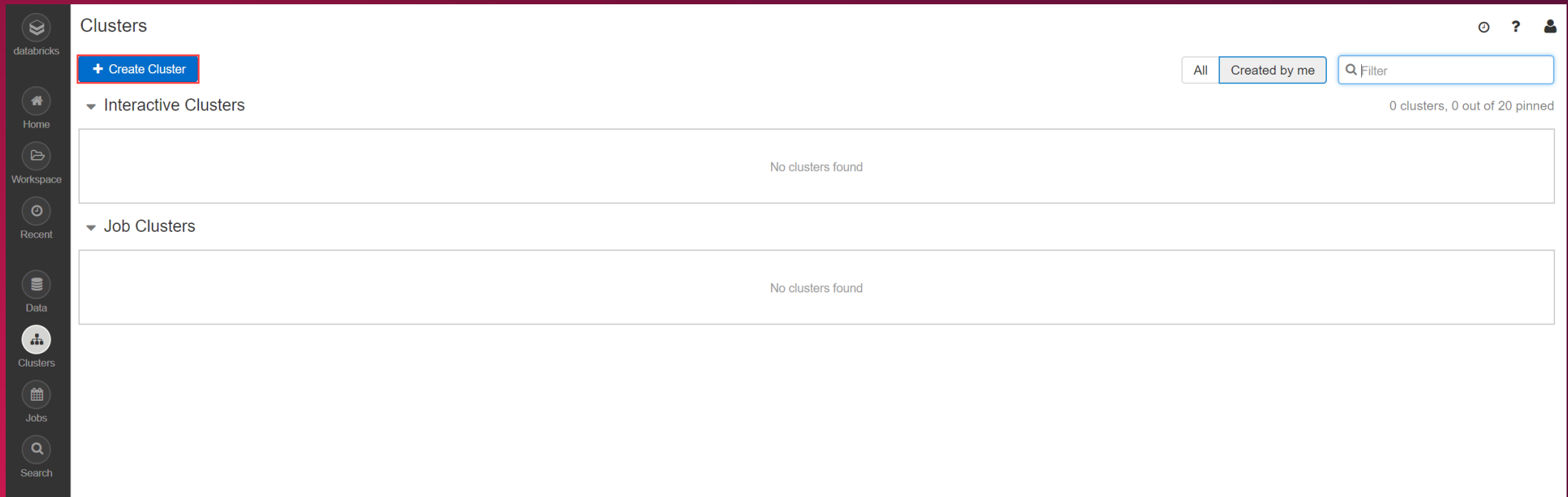
Práctica Apache Spark

Creando nuestro clúster en databricks

The screenshot displays the Databricks web interface. On the left, a dark sidebar contains navigation icons: Home, Workspace, Recent, Data, Clusters (circled in red with a red arrow pointing to it), Jobs, and Search. The main content area is white and features a 'Welcome to databricks' header. Below this, there are 'Featured Notebooks' with icons for Python and links to 'Introduction to Apache Spark on Databricks', 'Databricks for Data Scientists', and 'Introduction to Structured Streaming'. The interface is divided into three columns: 'New' (containing links to Notebook, Job, Cluster, Table, and Library), 'Documentation' (containing links to Databricks Guide, Python, R, Scala, SQL, and Importing Data), and 'What's new?' (listing 'Trash folder' and 'Cluster log retained for 30 days', with a link to 'Latest release notes'). The top right corner includes an 'Upgrade' button, a question mark, a user icon, and the text 'Community Edition (2.73.287)'.

Práctica Apache Spark

Creando nuestro clúster en databricks



The screenshot displays the Databricks Clusters management interface. On the left is a dark sidebar with navigation icons and labels: databricks, Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main content area is titled 'Clusters' and includes a '+ Create Cluster' button highlighted with a red box. To the right of the button are tabs for 'All' and 'Created by me', and a search bar labeled 'Filter'. Below these are two expandable sections: 'Interactive Clusters' and 'Job Clusters', both showing 'No clusters found'. A status message at the top right indicates '0 clusters, 0 out of 20 pinned'. The interface also features standard utility icons (refresh, help, user) in the top right corner.

Práctica Apache Spark

Creando nuestro clúster en databricks

Create Cluster

New Cluster Cancel Create Cluster **0 Workers:** 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU

Cluster Name
bda

Databricks Runtime Version ?
Runtime: 6.2 (Scala 2.11, Spark 2.4.4) | v

New This Runtime version supports only Python 3.

Instance

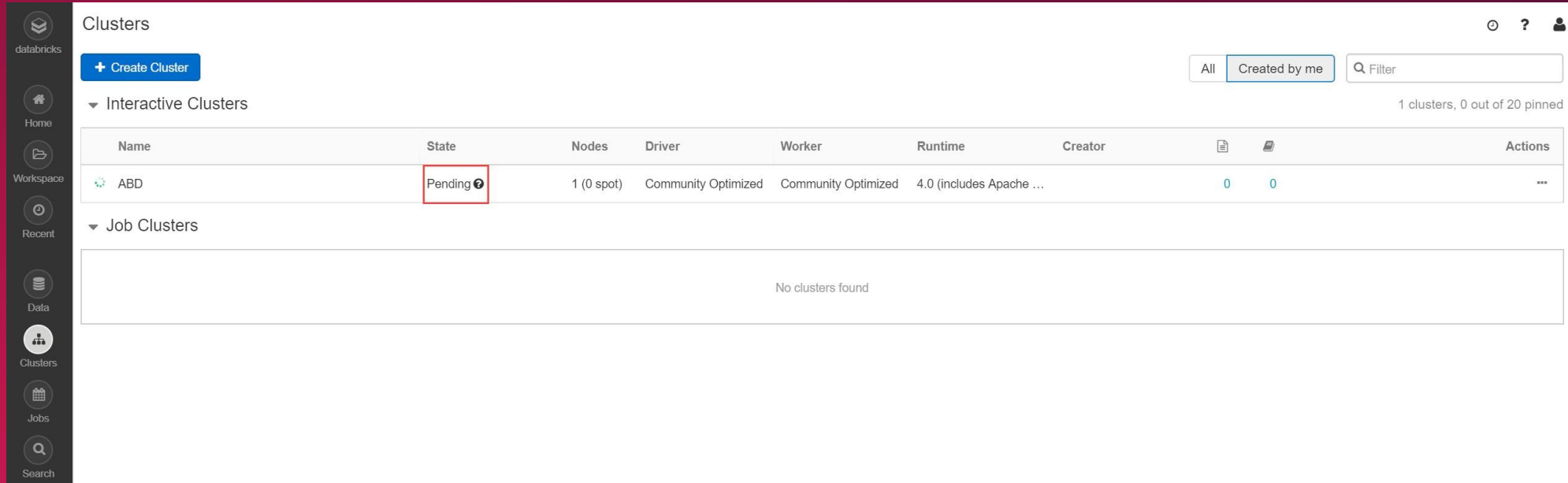
Free 6GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For [more configuration options](#), please [upgrade your Databricks subscription](#).

Instances [Spark](#)

Availability Zone ?
us-west-2c | v

Práctica Apache Spark

Creando nuestro clúster en databricks



The screenshot shows the Databricks Clusters management interface. On the left is a sidebar with navigation icons for Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main header area includes a '+ Create Cluster' button, a dropdown for 'Interactive Clusters', and a search filter. Below this, a table lists the cluster 'ABD' with a 'Pending' status, which is highlighted by a red box. The table also shows details for nodes, driver, worker, and runtime. Below the table, a section for 'Job Clusters' indicates that no clusters were found.

Clusters

+ Create Cluster

▼ Interactive Clusters 1 clusters, 0 out of 20 pinned


Name	State	Nodes	Driver	Worker	Runtime	Creator			Actions
ABD	Pending ⓘ	1 (0 spot)	Community Optimized	Community Optimized	4.0 (includes Apache ...		0	0	...


▼ Job Clusters


No clusters found


Práctica Apache Spark


Creando nuestro clúster en databricks



databricks



Home



Workspace


Recents


Data


Clusters


Jobs


Search

Clusters


[+ Create Cluster](#)

AllCreated by me

Filter

1 clusters, 0 pinned

▼ Interactive Clusters

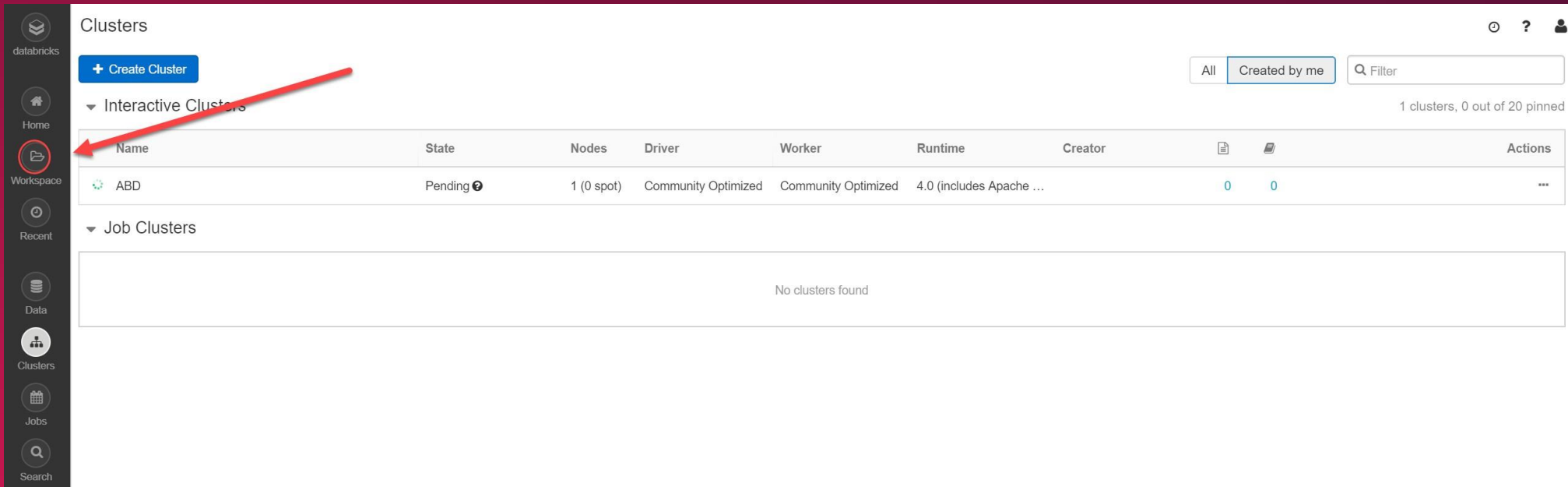
Name	State	Nodes	Driver	Worker	Runtime	Creator		Actions
 bda	Running	1 (0 spot)	Community Optimi..	Community Optimi..	6.2 (includes Apach..	cindylopez.sw@g...	0	...

▼ Automated Clusters

No clusters found

Práctica Apache Spark

Creando nuestro Notebook



The screenshot shows the Databricks Clusters management interface. On the left sidebar, the 'Workspace' icon is highlighted with a red circle and a red arrow points to it. The main content area is titled 'Clusters' and features a '+ Create Cluster' button. Below this, there are two sections: 'Interactive Clusters' and 'Job Clusters'. The 'Interactive Clusters' section contains a table with one cluster named 'ABD' in a 'Pending' state. The 'Job Clusters' section is currently empty, displaying 'No clusters found'.

Clusters

[+ Create Cluster](#)

▼ Interactive Clusters 1 clusters, 0 out of 20 pinned

Name	State	Nodes	Driver	Worker	Runtime	Creator			Actions
ABD	Pending ⓘ	1 (0 spot)	Community Optimized	Community Optimized	4.0 (includes Apache ...		0	0	...

▼ Job Clusters

No clusters found

Práctica Apache Spark

Creando nuestro Notebook

The screenshot shows the Databricks web interface. On the left sidebar, the 'Users' menu item is highlighted. In the top navigation bar, the 'Clusters' tab is selected. A 'Create' dropdown menu is open, showing options: 'Notebook' (highlighted), 'Library', 'Folder', 'Export', 'Import', 'Clone', and 'Permissions'. Below the navigation bar, there is a table of clusters. The table has columns: State, Nodes, Driver, Worker, and Runtime. One cluster is listed: 'ABD' with state 'Running', 1 (0 spot) nodes, and runtime '4.0 (includes Apache)'. Below this, a section for 'Job Clusters' shows 'No clusters found'.

State	Nodes	Driver	Worker	Runtime
Running	1 (0 spot)	Community Optimized	Community Optimized	4.0 (includes Apache)

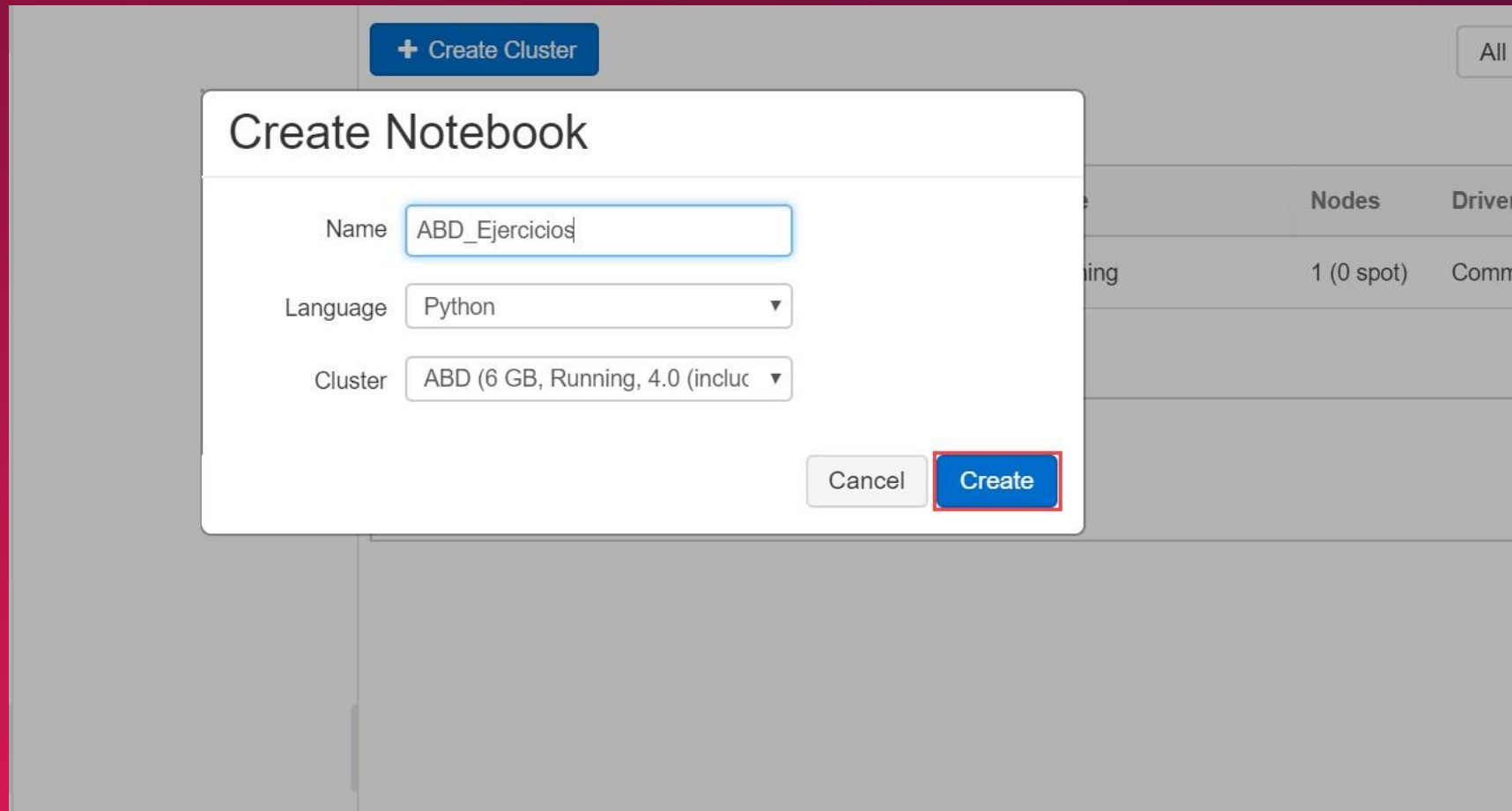
1 clusters, 0 out of 20 pinned

▼ Job Clusters

No clusters found

Práctica Apache Spark

Creando nuestro Notebook



The image shows a 'Create Notebook' dialog box overlaid on a blurred background of a cloud management console. The dialog box has a title bar 'Create Notebook' and three input fields: 'Name' with the value 'ABD_Ejercicios', 'Language' with a dropdown menu showing 'Python', and 'Cluster' with a dropdown menu showing 'ABD (6 GB, Running, 4.0 (includ'. At the bottom right of the dialog are two buttons: 'Cancel' and 'Create'. The 'Create' button is highlighted with a red border. In the background, a table with columns 'Nodes' and 'Driver' is partially visible, showing a row with '1 (0 spot)' and 'Comm'.

Create Notebook

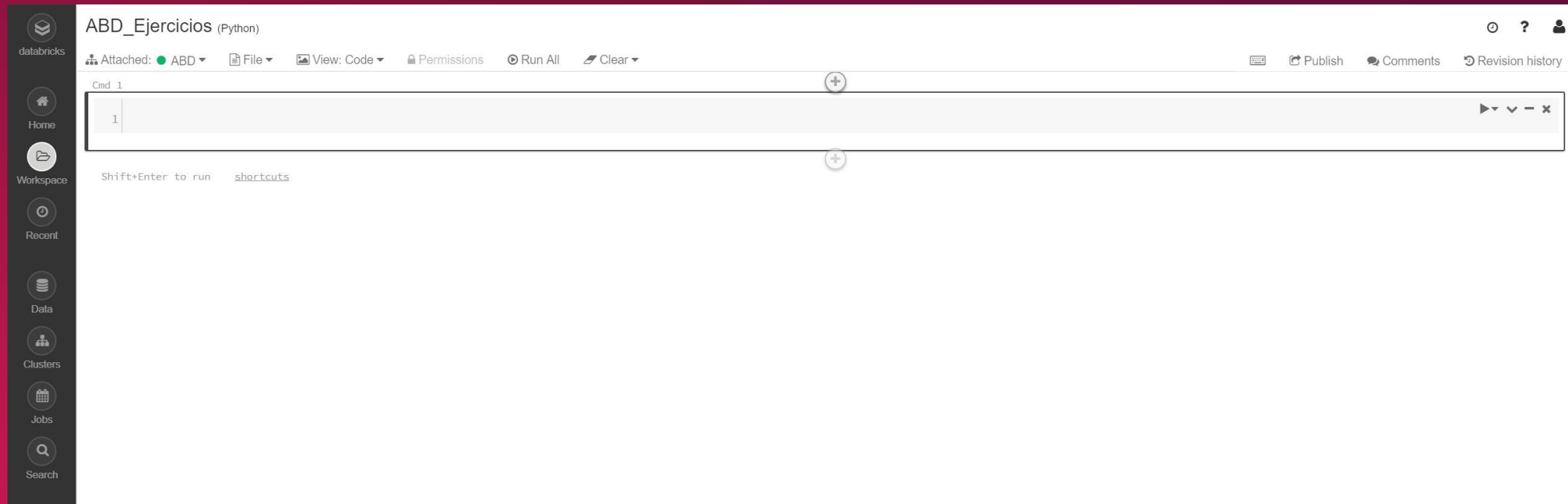
Name

Language

Cluster

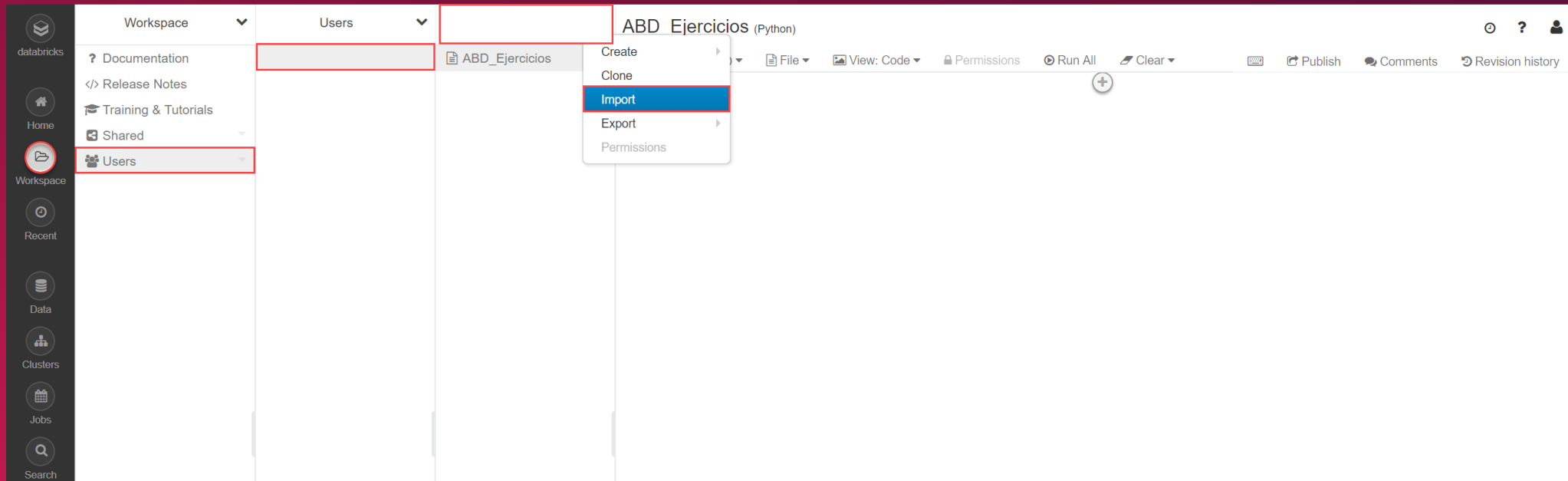
Práctica Apache Spark

Creando nuestro Notebook



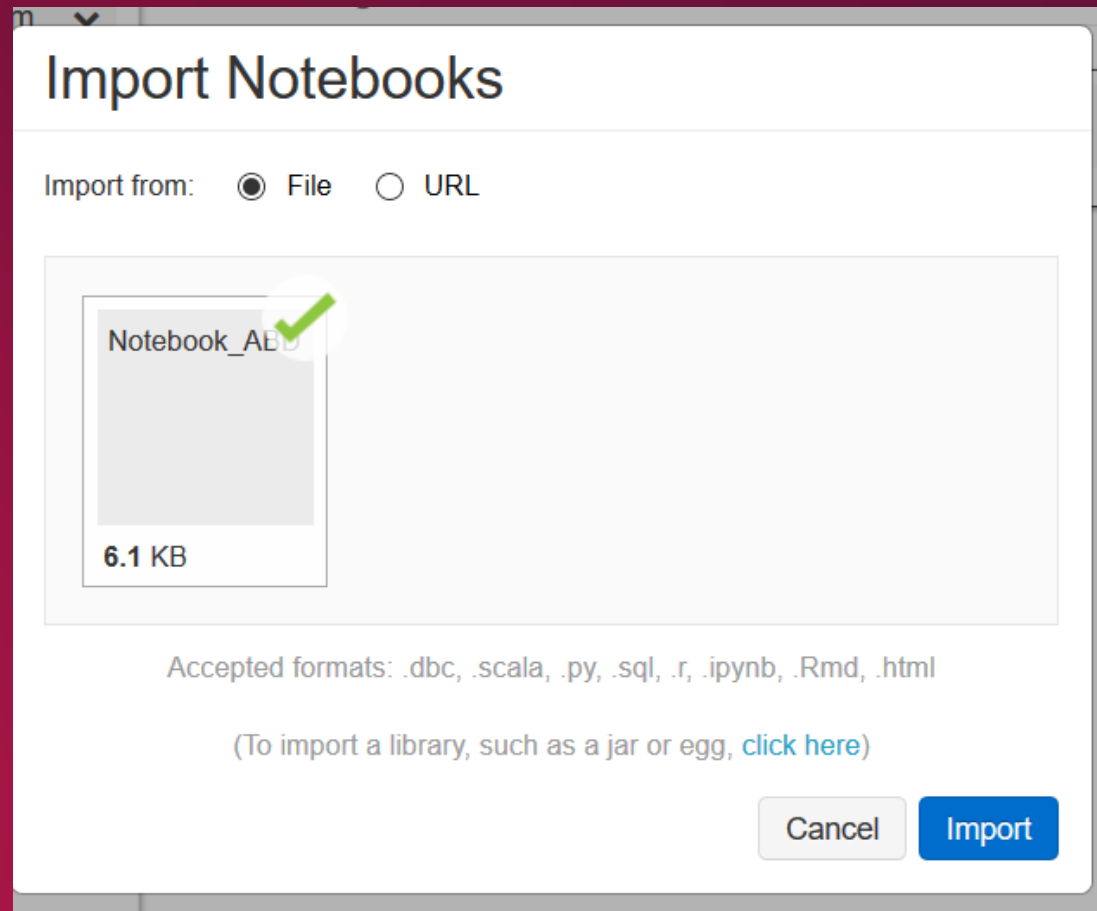
Práctica Apache Spark

Importando datos a nuestro Notebook



Práctica Apache Spark

Importando un Notebook



Práctica Apache Spark

Importando un Notebook

Notebook_ABD (Python)

Detached File View: Code Permissions Run All Clear Publish Comments Revision history

Cmd 1

Clase 19. Práctica de Apache Spark. Curso Análisis de BIG DATA.

Este cuaderno se basa en tutoriales realizados por Databricks. Databricks is a leading provider of the commercial and enterprise supported version of Spark.

[Databricks](#)

Cmd 2

```
1 # Section 1A
2 # This is a Python cell. You can run normal Python code here...
3 print 'The sum of 1 and 1 is {}'.format(1+1)
4
5 # Here is another Python cell, this time with a variable (x) declaration and an if statement:
6 x = 42
7 if x > 40:
8     print 'The sum of 1 and 2 is {}'.format(1+2)
```

Cmd 3

```
1 # Section 1B
2
3 # This cell relies on x being defined already.
4 # If we didn't run the cells from part (1a) this code would fail.
5 print x * 2
```

Cmd 4

```
1 # Section 1C
2
3 # Import the regular expression library
4 import re
5 m = re.search('(<=abc)def', 'abcdef')
6 m.group(0)
7
8 # Import the datetime library
9 import datetime
10 print 'This was last run on: {}'.format(datetime.datetime.now())
11
```

Send Feedback

Práctica Apache Spark

Ejecutando nuestro Notebook

The screenshot displays the Databricks Notebook interface. On the left is a dark sidebar with navigation icons for Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main area is titled 'Notebook_ABD (Python)' and includes a toolbar with options like 'Detached', 'File', 'View: Code', 'Permissions', 'Run All', and 'Clear'. Below the toolbar, a cluster selection dropdown shows 'ABD (6 GB, Running, 4.0 (includes Apache Spark 2.3.0, Scala 2.11))' selected. The notebook content begins with a header 'Clase 19. Práctica de Apache Spark. Curso Análisis de BIG DATA.' followed by an introductory paragraph about Databricks. The notebook contains three code cells, each labeled 'Cmd' followed by a number. The first cell (Cmd 2) contains Python code for a simple sum calculation. The second cell (Cmd 3) contains code that relies on a variable defined in the first cell. The third cell (Cmd 4) contains code for importing libraries and using regular expressions and datetime. A 'Send Feedback' button is located at the bottom right of the interface.

Notebook_ABD (Python)

Attach to:
ABD (6 GB, Running, 4.0 (includes Apache Spark 2.3.0, Scala 2.11))

Clase 19. Práctica de Apache Spark. Curso Análisis de BIG DATA.

Este cuaderno se basa en tutoriales realizados por Databricks. Databricks is a leading provider of the commercial and enterprise supported version of Spark.

[Databricks](#)

Cmd 2

```
1 # Section 1A
2 # This is a Python cell. You can run normal Python code here...
3 print 'The sum of 1 and 1 is {}'.format(1+1)
4
5 # Here is another Python cell, this time with a variable (x) declaration and an if statement:
6 x = 42
7 if x > 40:
8     print 'The sum of 1 and 2 is {}'.format(1+2)
```

Cmd 3

```
1 # Section 1B
2
3 # This cell relies on x being defined already.
4 # If we didn't run the cells from part (1a) this code would fail.
5 print x * 2
```

Cmd 4

```
1 # Section 1C
2
3 # Import the regular expression library
4 import re
5 m = re.search('(?<=abc)def', 'abcdef')
6 m.group(0)
7
8 # Import the datetime library
9 import datetime
10 print 'This was last run on: {}'.format(datetime.datetime.now())
11
```

Send Feedback

Práctica Apache Spark

Ejecutando nuestro Notebook

Notebook_ABD (Python)

Attached: ABD File View: Code Permissions Run All Clear

Cmd 1

Clase 19. Práctica de Apache Spark. Curso Análisis de BIG DATA.

Este cuaderno se basa en tutoriales realizados por Databricks. Databricks is a leading provider of the commercial and enterprise supported version of Spark.

[Databricks](#)

Cmd 2

```
1 # Section 1A
2 # This is a Python cell. You can run normal Python code here...
3 print 'The sum of 1 and 1 is {}'.format(1+1)
4
5 # Here is another Python cell, this time with a variable (x) declaration and an if statement:
6 x = 42
7 if x > 40:
8     print 'The sum of 1 and 2 is {}'.format(1+2)
```

Cmd 3

```
1 # Section 1B
2
3 # This cell relies on x being defined already.
4 # If we didn't run the cells from part (1a) this code would fail.
5 print x * 2
```

Cmd 4

```
1 # Section 1C
2
3 # Import the regular expression library
4 import re
5 m = re.search('(<=abc)def', 'abcdef')
6 m.group(0)
7
8 # Import the datetime library
9 import datetime
10 print 'This was last run on: {}'.format(datetime.datetime.now())
11
```

Send Feedback

Práctica Apache Spark

Revisando los resultados

Home

Workspace

Recent

Data

Clusters

Jobs

Search

Notebook_ABD (Python)

Attached: ABD File View: Results Only Permissions Run All Clear

Revision history

Clase 19. Práctica de Apache Spark. Curso Análisis de BIG DATA.

Este cuaderno se basa en tutoriales realizados por Databricks. Databricks is a leading provider of the commercial and enterprise supported version of Spark.

[Databricks](#)

The sum of 1 and 1 is 2
The sum of 1 and 2 is 3

84

This was last run on: 2018-06-15 18:07:32.083844

Out[4]: `__main__.RemoteContext`

```
Out[5]:
['PACKAGE_EXTENSIONS',
 '__class__',
 '__delattr__',
 '__dict__',
 '__doc__',
 '__enter__',
 '__exit__',
 '__format__',
 '__getattr__',
 '__getnewargs__',
 '__hash__',
 '__init__',
 '__module__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
```

Help on RemoteContext in module __main__ object:

[Send Feedback](#)

Ejercicio – Analizar y documentar los ejercicios del notebook