

## Tutorial Databricks

<b>Disciplina</b>	<b>FAM – Fundamentos de Aprendizado de Máquina</b>
-------------------	--

### Objetivos

Exercitar os seguintes conceitos vistos em sala de aula:

- ✓ Transformações;
- ✓ Consultas;
- ✓ Leitura e escrita de dados.

### Enunciado

Este tutorial consiste em apresentar a vocês um dos sites mais utilizados pelos cientistas de dados/Engenheiros de Machine Learning que desejam empregar o Spark em suas aplicações e é, certamente, uma excelente plataforma de aprendizado e compartilhamento de dados. Este é o site databricks.com. Neste tutorial, será mostrado como acessar, criar uma conta e utilizar a plataforma para rodar os seus programas utilizando o PySpark. Essa ferramenta poderá ser utilizada para o desenvolvimento do trabalho prático.

### Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Seguir o tutorial de utilização do Databricks;
2. Compilar um programa simples no “Notebook” do Databricks;

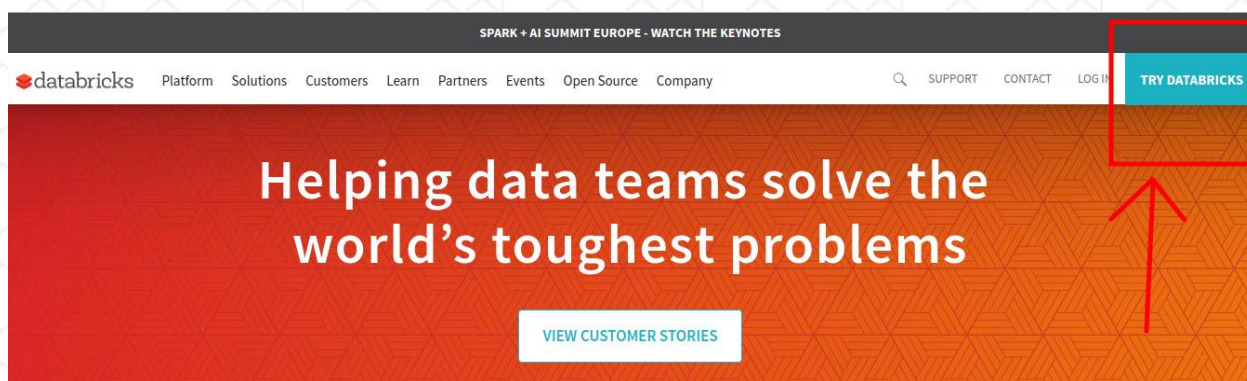
### Tutorial

Neste tutorial, vamos aprender a utilizar o DataBricks. Databricks é um ambiente online para o aprendizado e desenvolvimento de aplicações que utilizam sistemas de processamento distribuído. Databricks pode ser completamente integrado a plataformas como Amazon AWS e Microsoft Azure.

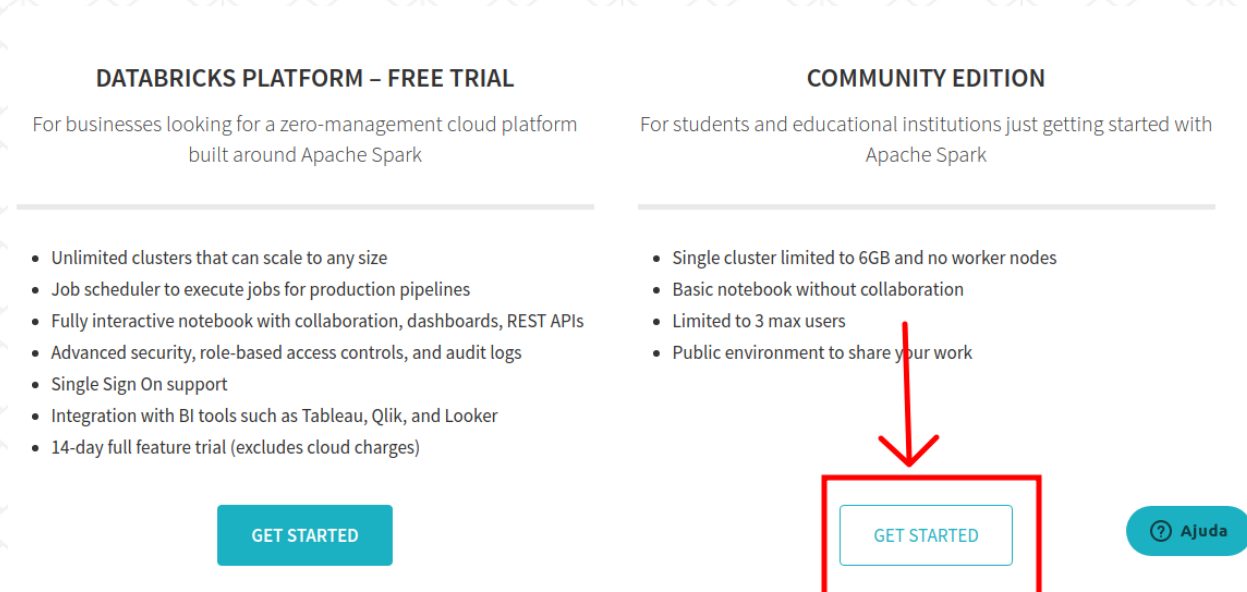
Para acessar o ambiente de projeto gratuito é necessário criar uma conta no Databricks Community. Acesse o link:

<https://databricks.com/>

Clique em “Try Databricks”, como mostra a figura abaixo.



Após esse passo, você verá uma tela com duas opções para criação da conta. Escolha a opção “Community Edition” e clique no botão “Get Started”, conforme mostra a figura abaixo:



Na próxima tela, preencha os campos com os seus dados e clique em “**Sign Up**”. Preencha os campos marcados conforme as opções apresentadas na figura abaixo:

## Sign Up for Databricks Community Edition

First Name \*

Last Name \*

Company Name \*

Work Email \*

Phone Number

What is your intended use case? \*


Personal - Learning Spark


How would you describe your role? \*

Student

☐ Keep me informed with the occasional update about Databricks and Apache Spark™.

By clicking "Sign Up", you agree to the [Terms of Service](#) and the [Privacy Policy](#).

☐ Não sou um robô  reCAPTCHA  
Privacidade - Termos



**Sign Up**

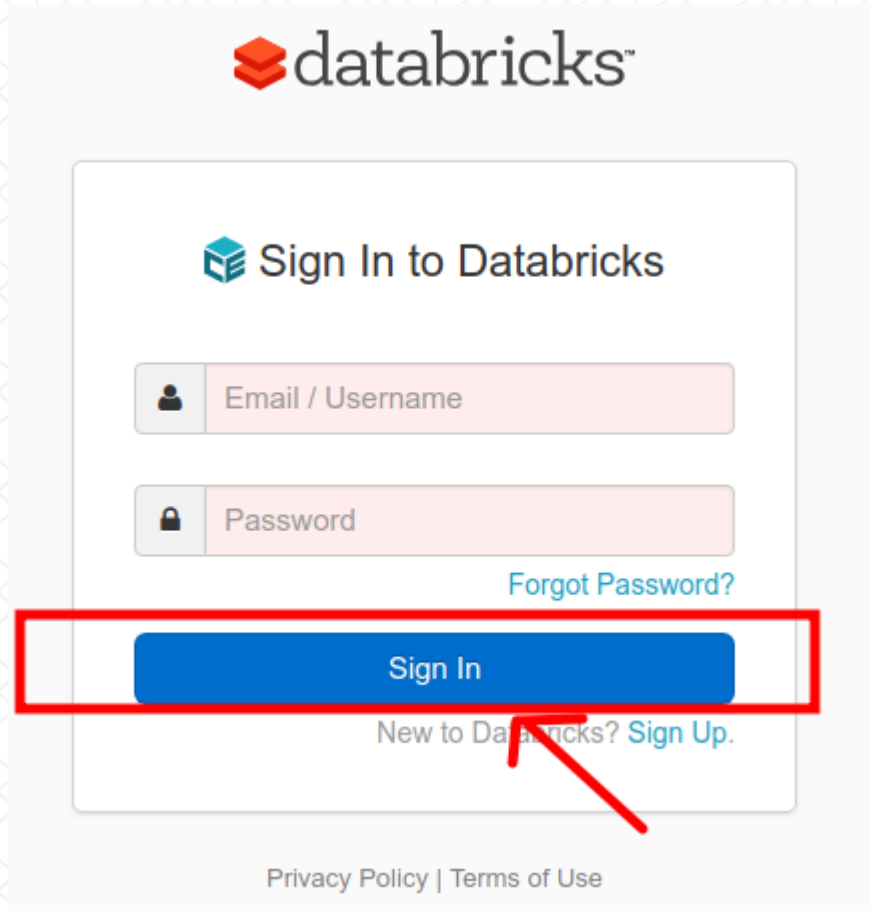
**OBS:** Caso não consiga realizar o cadastro (o botão de “Sign Up” não possa ser acionado), tente esse processo em outro navegador (Mozilla Firefox, Opera ou Windows Explorer). Às vezes o cadastramento apresenta incompatibilidade com o Google Chrome.

Após realizar esse processo, você receberá um e-mail de confirmação. Acesse o endereço de e-mail cadastrado e confirme a mensagem.

Para acessar o **Databricks Community**, digite o endereço abaixo em seu navegador:

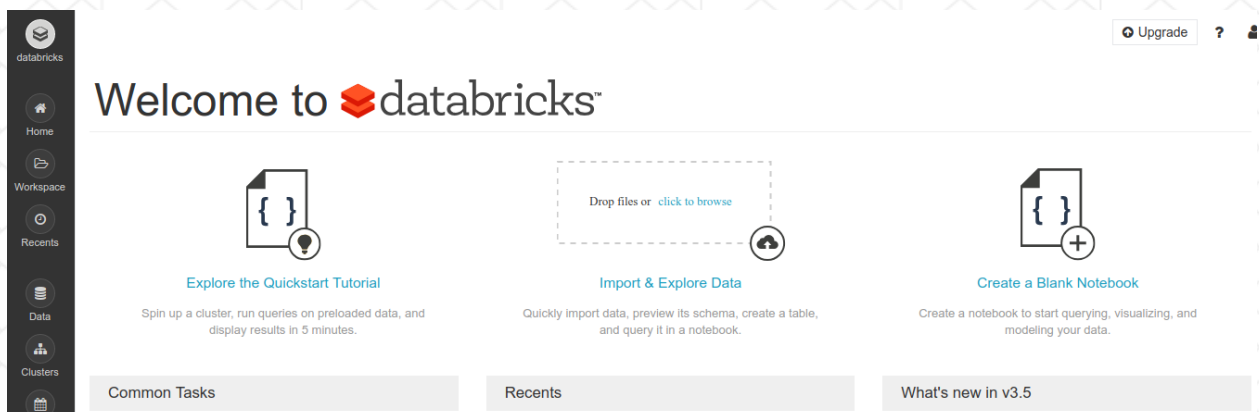


<https://community.cloud.databricks.com/login.html>



The image shows the Databricks login interface. At the top is the Databricks logo. Below it is a box titled "Sign In to Databricks". Inside this box are two input fields: "Email / Username" and "Password". To the right of the password field is a link "Forgot Password?". Below the input fields is a large blue "Sign In" button, which is highlighted with a red rectangle. Below the button is a link "New to Databricks? Sign Up.". At the bottom of the box are links for "Privacy Policy" and "Terms of Use".

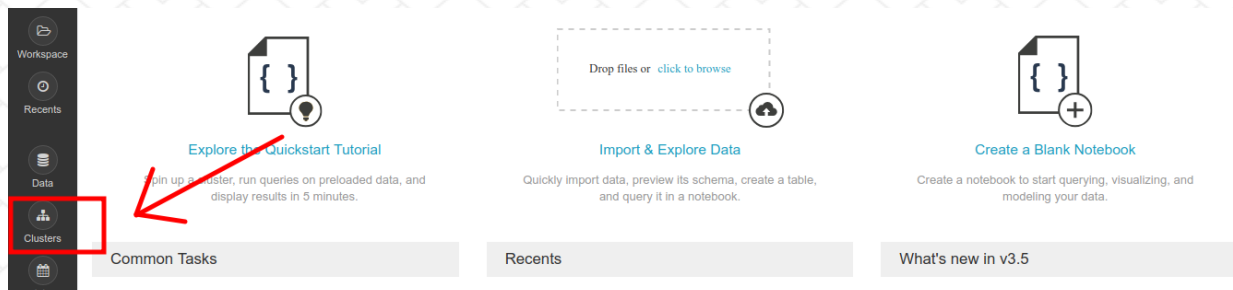
Adicione o seu login/email e senha cadastradas, assim terá acesso ao ambiente como mostrado abaixo:



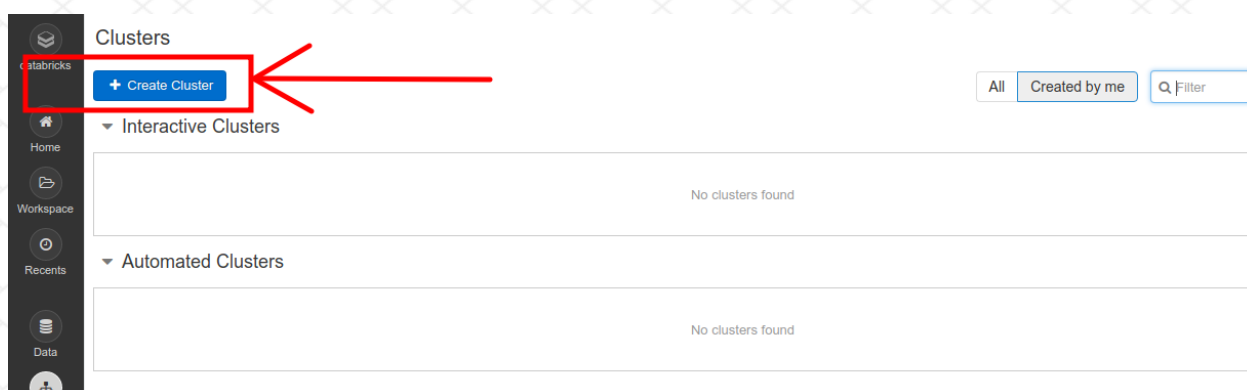
The image shows the Databricks "Welcome to databricks" screen. On the left is a vertical sidebar with icons for Home, Workspace, Recents, Data, and Clusters. The main area has a header "Welcome to databricks" and three main action cards: "Explore the Quickstart Tutorial", "Import & Explore Data", and "Create a Blank Notebook". Each card has a brief description of what it does. At the bottom are three tabs: "Common Tasks", "Recents", and "What's new in v3.5".

**Criando uma primeira aplicação:**

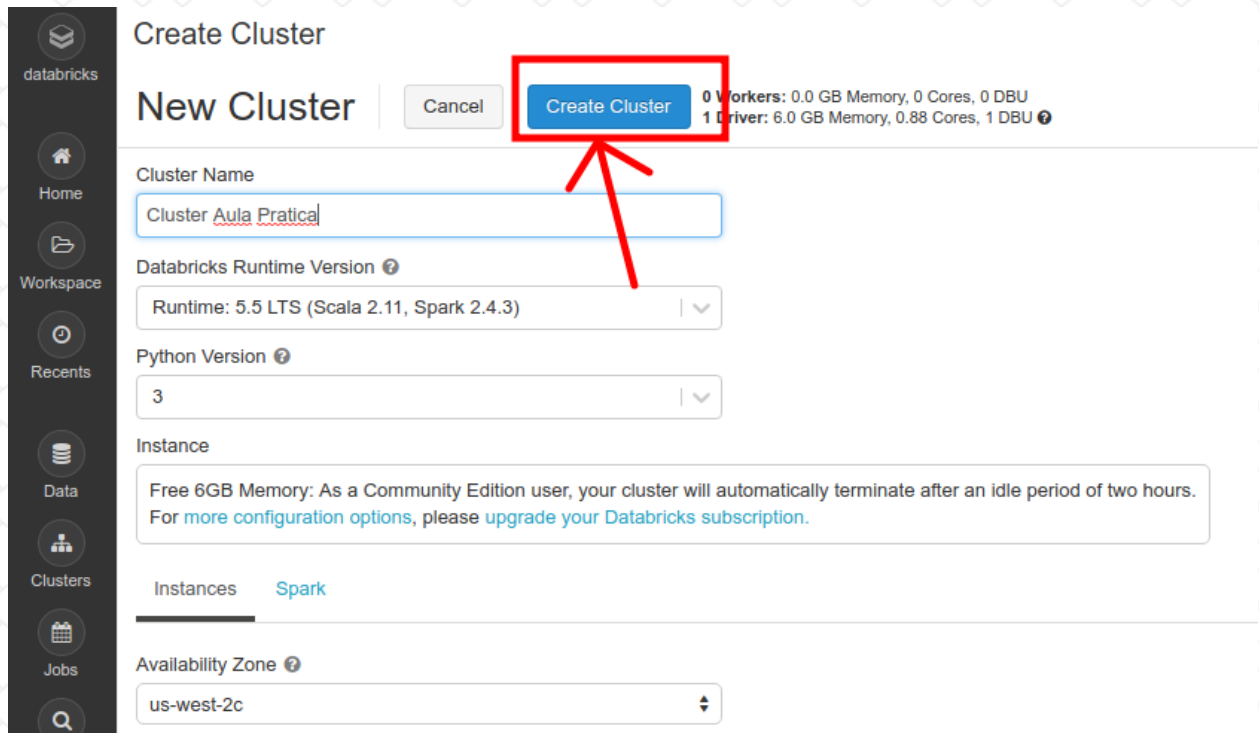
Para iniciar o processo de criação de aplicações é necessário realizar a configuração do Cluster. Para isso, no menu lateral esquerdo, acesse a aba “**Cluster**”, conforme mostra a figura abaixo:



Ao acessar esse menu, você será direcionado para uma nova página. Essa página é onde vamos configurar o “**Cluster**”. Para iniciar esse processo, clique em “**Create Cluster**”, conforme mostra a figura abaixo:



Na próxima página, digite um nome para o Cluster e clique em “**Create Cluster**”. Conforme mostra a figura abaixo:



**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  
Cluster Aula Pratica

Databricks Runtime Version ?  
Runtime: 5.5 LTS (Scala 2.11, Spark 2.4.3) | v

Python Version ?  
3 | v

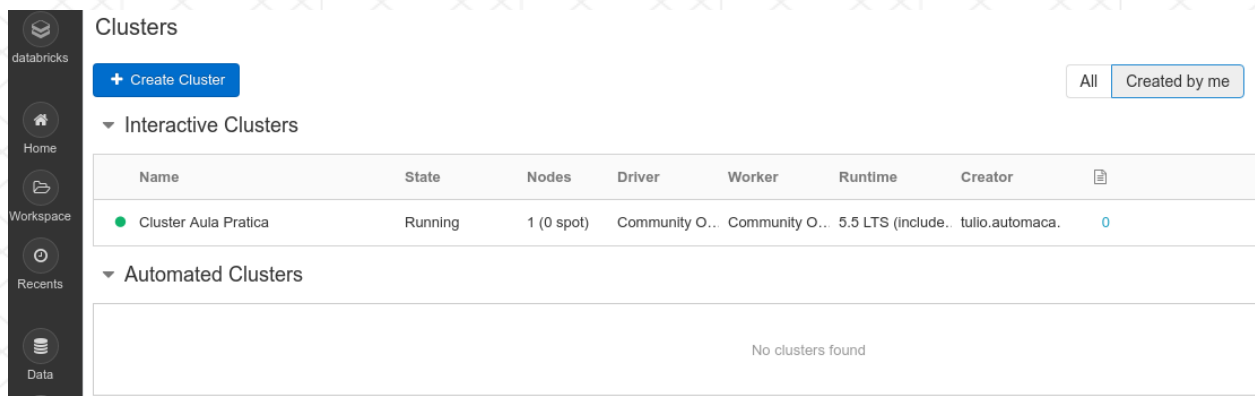
Instance

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

Após esse processo, será criado o cluster com o nome inserido. A figura abaixo apresenta o resultado final presente na aba “**Cluster**”.



**Clusters**

+ Create Cluster All Created by me

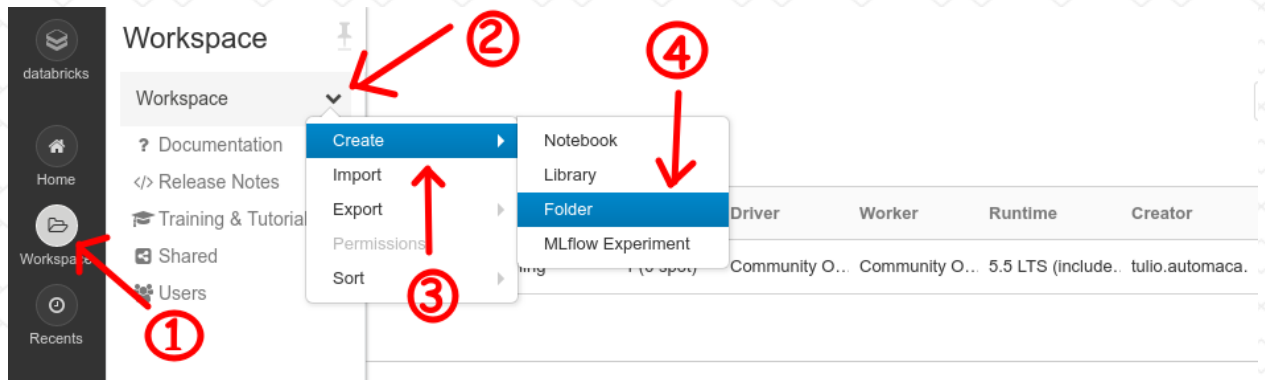
▼ Interactive Clusters

Name	State	Nodes	Driver	Worker	Runtime	Creator	
Cluster Aula Pratica	Running	1 (0 spot)	Community O...	Community O...	5.5 LTS (include..	tulio.automaca.	0

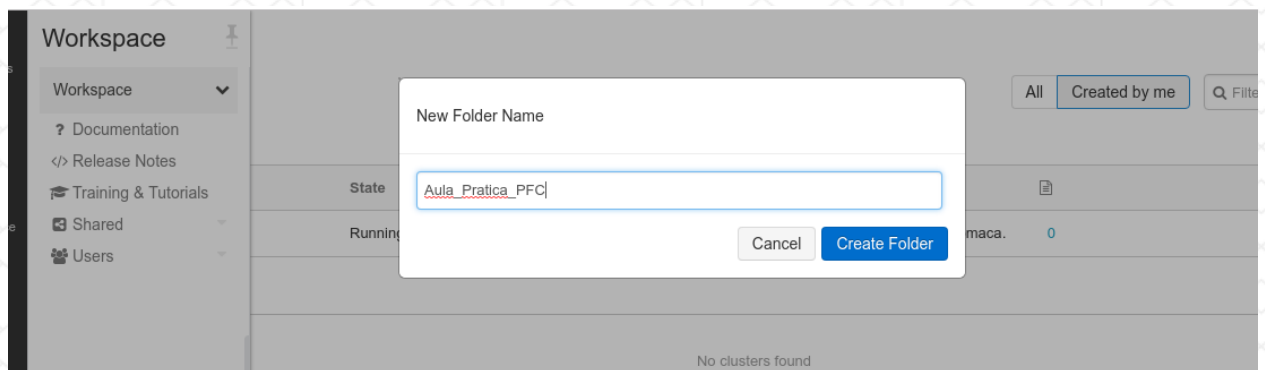
▼ Automated Clusters

No clusters found

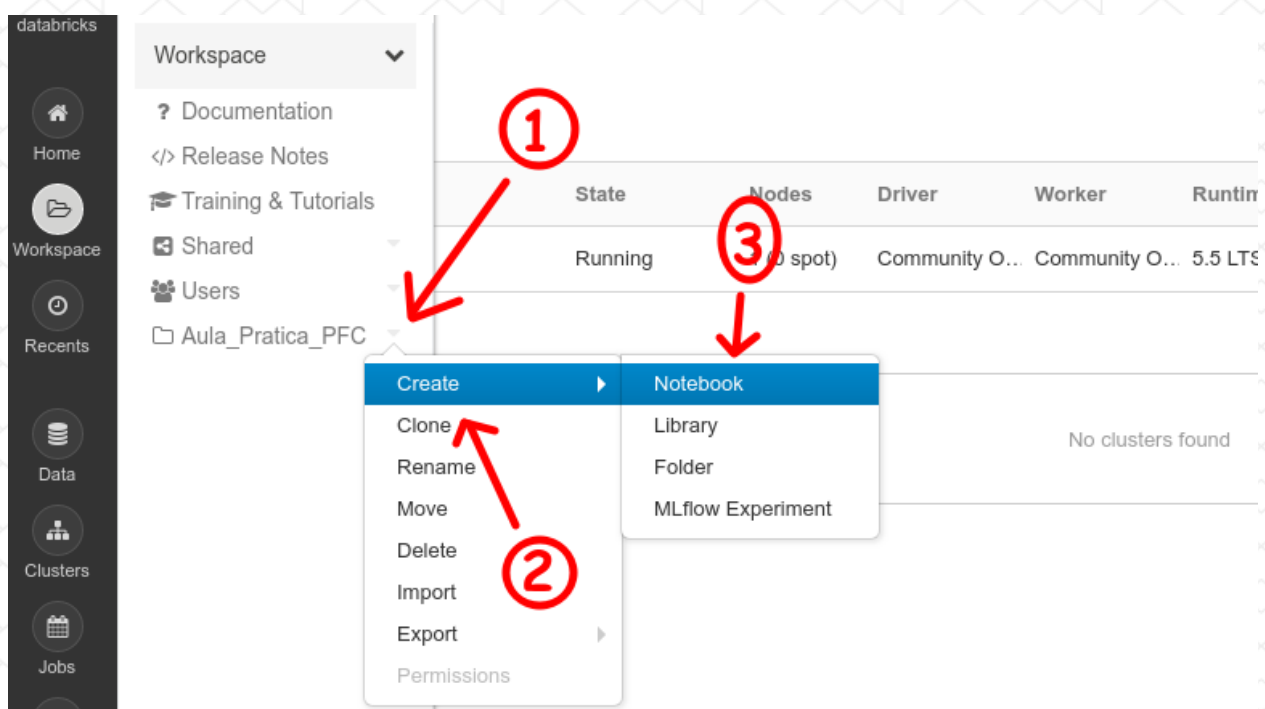
Vamos criar a nossa primeira aplicação utilizando o Databricks. Para isso, acesse, no menu lateral, a aba “**Workspace**”. Nessa aba, clique na seta superior direita, conforme mostra a figura abaixo, selecione “**Create**” e, por último, selecione “**Folder**”. A figura abaixo mostra as etapas a serem seguidas.



Na próxima aba, adicione um nome para essa pasta.

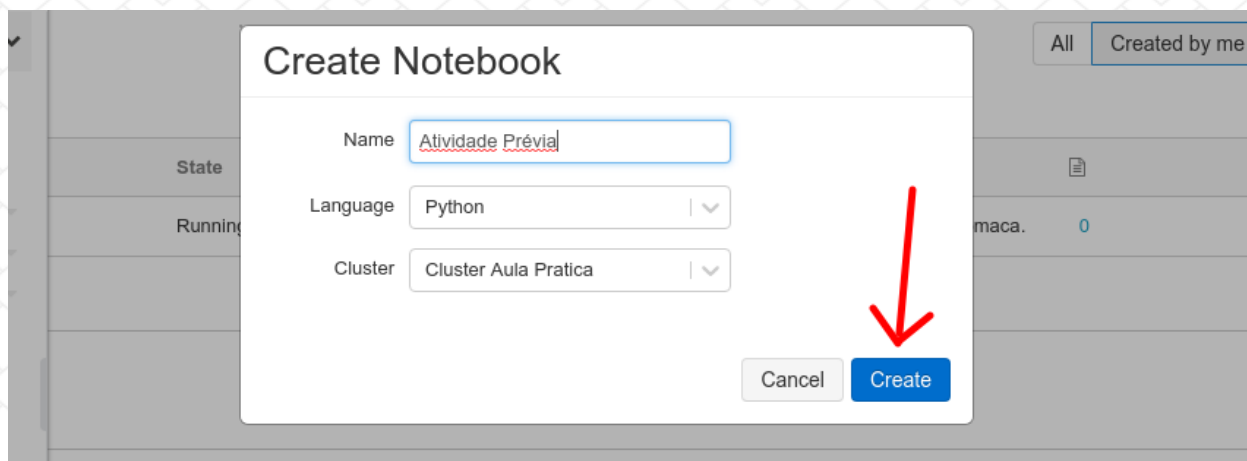


Com a pasta criada, clique na seta em frente ao nome dessa pasta. Selecione "Create" e depois "Notebook". A figura abaixo apresenta a sequência de passos descrita.

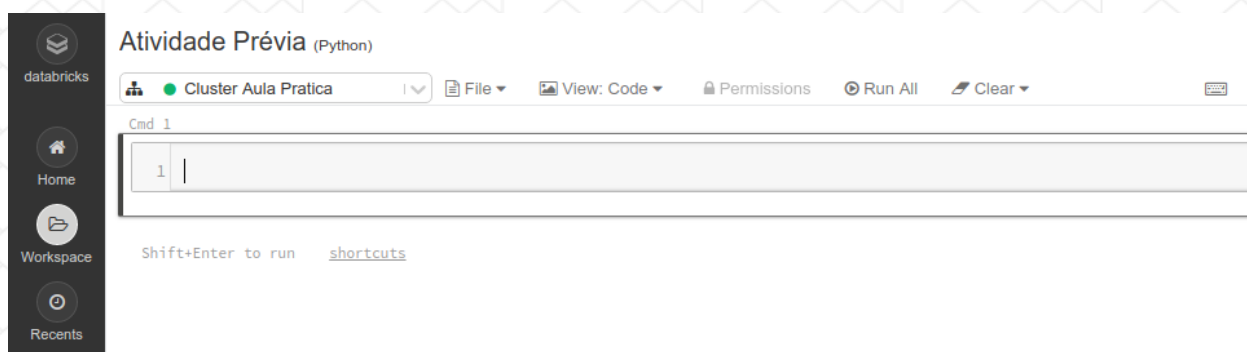




Adicione um nome para o **"Notebook"**. Não será preciso alterar nenhum parâmetro. Apenas confira se os dados estão como mostrados na figura abaixo e clique em **"Create"**.



Pronto, temos a configuração do ambiente para a realização da nossa atividade prática.



Agora, vamos rodar o nosso primeiro programa escrito em Python que utiliza o Databricks. Para isso, digite o comando abaixo na primeira célula do **"Notebook"**.

```
%python
from datetime import datetime

# Data atual
print(datetime.now())
```



Após esse processo, temos que executar a célula. Para isso, você pode pressionar “**Ctrl + Enter**” ou clicar no botão “**Play**” que aparece na lateral direita superior da célula. A figura abaixo mostra o resultado da execução e como acessar o “**Play**” da célula.

