

Bootcamp: Cientista de Dados

Tutorial Batabricks

Módulo 3: Técnicas para o processamento do Big Data

Objetivos de Ensino

Exercitar os seguintes conceitos vistos em sala de aula:

- Spark MLlib.

Enunciado

Esta Atividade Prévia consiste em apresentar a vocês um dos sites mais utilizados pelos cientistas de dados 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. Nesta atividade será apresentado um tutorial de como acessar, criar uma conta e utilizar a plataforma para rodar os seus programas utilizando o PySpark.

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.
3. Tirar um print da tela do PC com a saída do programa executado.

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. O 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:

DATABRICKS PLATFORM – FREE TRIAL	COMMUNITY EDITION
For businesses looking for a zero-management cloud platform built around Apache Spark	For students and educational institutions just getting started with Apache Spark
<ul style="list-style-type: none">• Unlimited clusters that can scale to any size• Job scheduler to execute jobs for production pipelines• Fully interactive notebook with collaboration, dashboards, REST APIs• Advanced security, role-based access controls, and audit logs• Single Sign On support• Integration with BI tools such as Tableau, Qlik, and Looker• 14-day full feature trial (excludes cloud charges)	<ul style="list-style-type: none">• Single cluster limited to 6GB and no worker nodes• Basic notebook without collaboration• Limited to 3 max users• Public environment to share your work
GET STARTED	GET STARTED Ajuda

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 

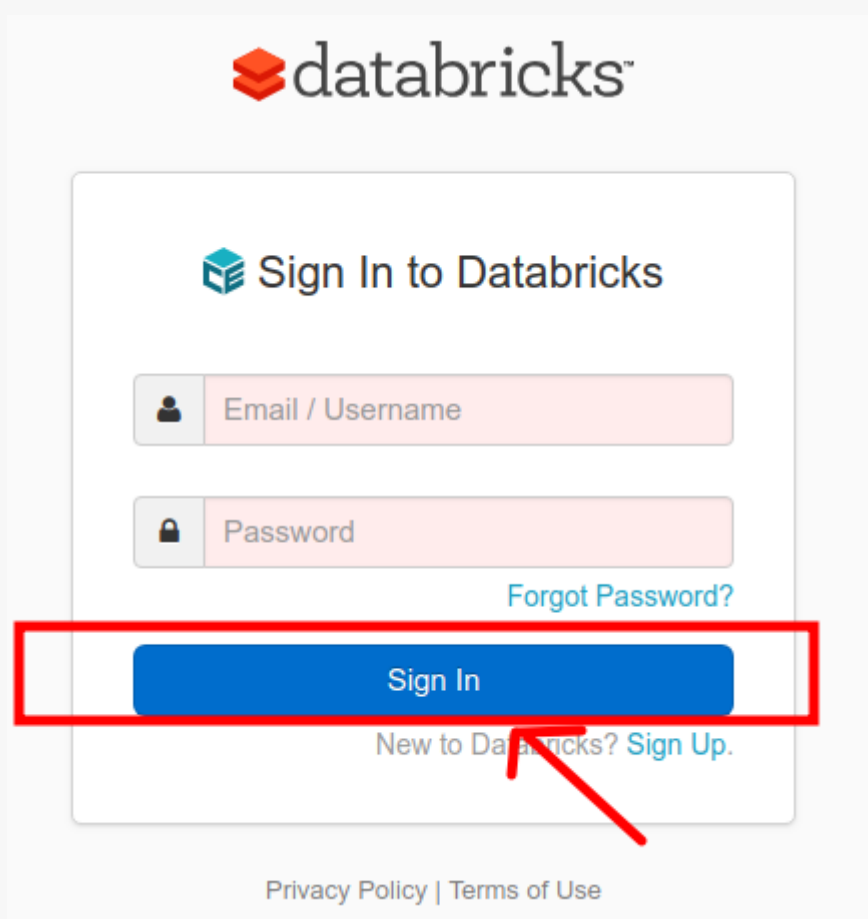
[Sign Up](#)

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>.



databricks™

Sign In to Databricks

Email / Username

Password

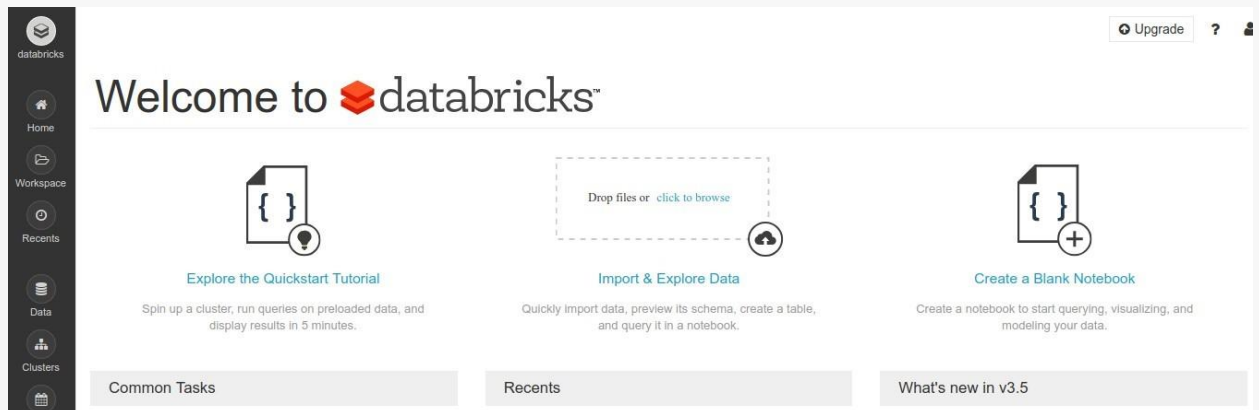
[Forgot Password?](#)

Sign In

New to Databricks? [Sign Up.](#)

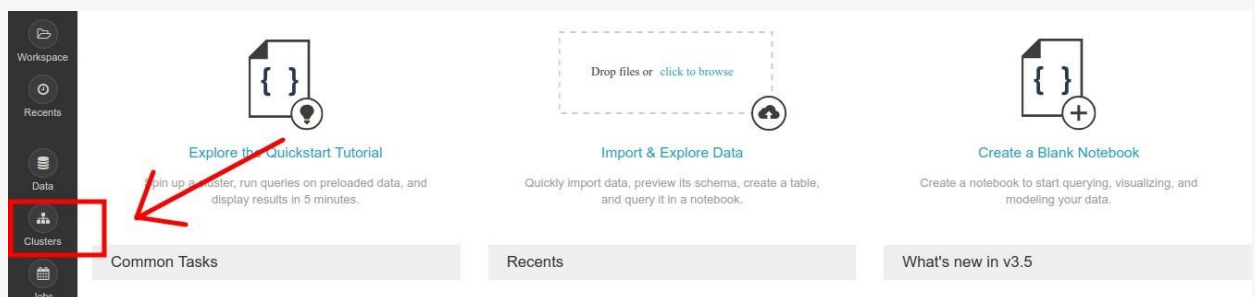
[Privacy Policy](#) | [Terms of Use](#)

Adicione o seu login/e-mail e senha cadastrado, assim terá acesso ao ambiente como mostrado abaixo:

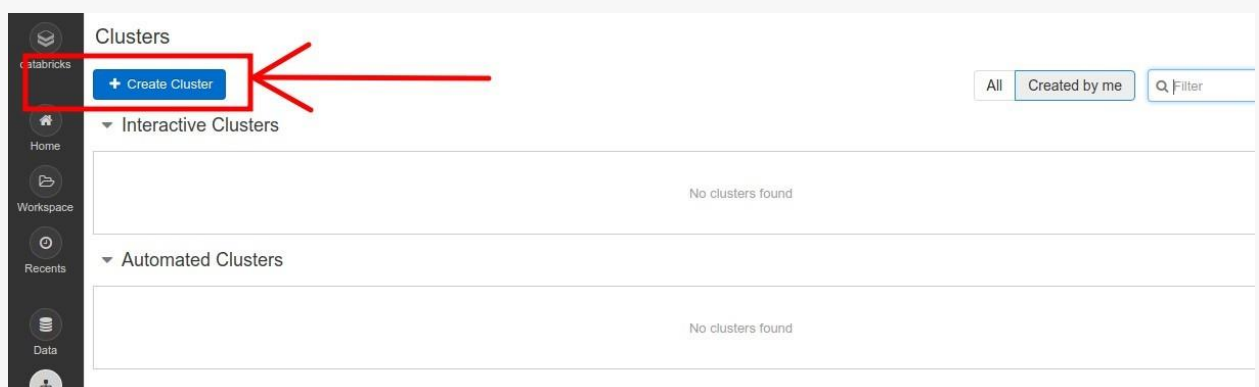


Criando uma primeira aplicação:

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

Databricks Runtime Version ?

Python Version ?

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 ?

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

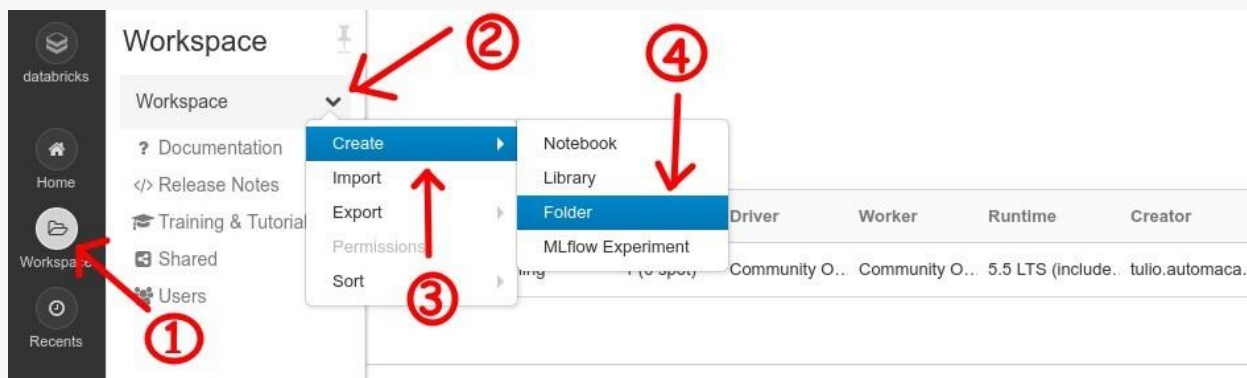
▼ Interactive Clusters

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

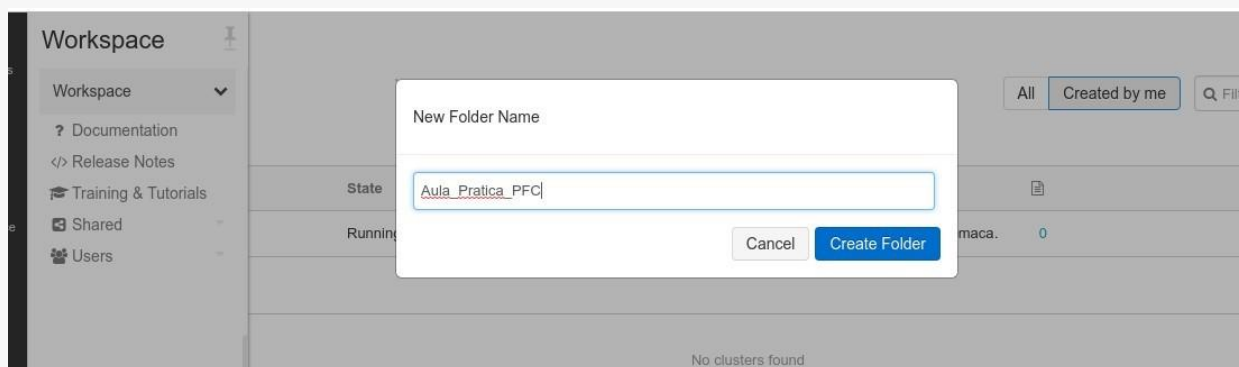
▼ Automated Clusters

No clusters found

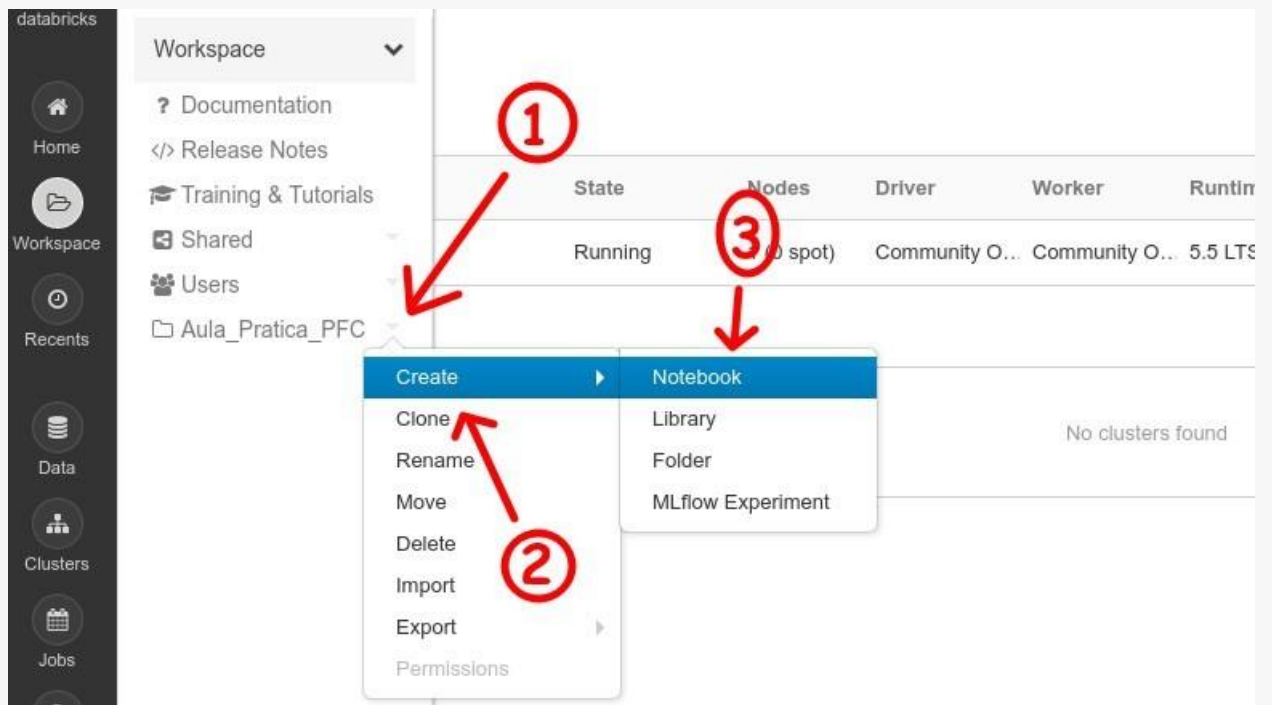
Vamos criar a nossa primeira aplicação utilizando o Databricks. Para isso, acesse, no menu lateral, a aba “Workspace”. Nessa aba clique na seta superior direita, conforme mostra a figura abaixo, selecione “Create” e, por último, selecione “Folder”. A figura abaixo mostra as etapas a serem seguidas.



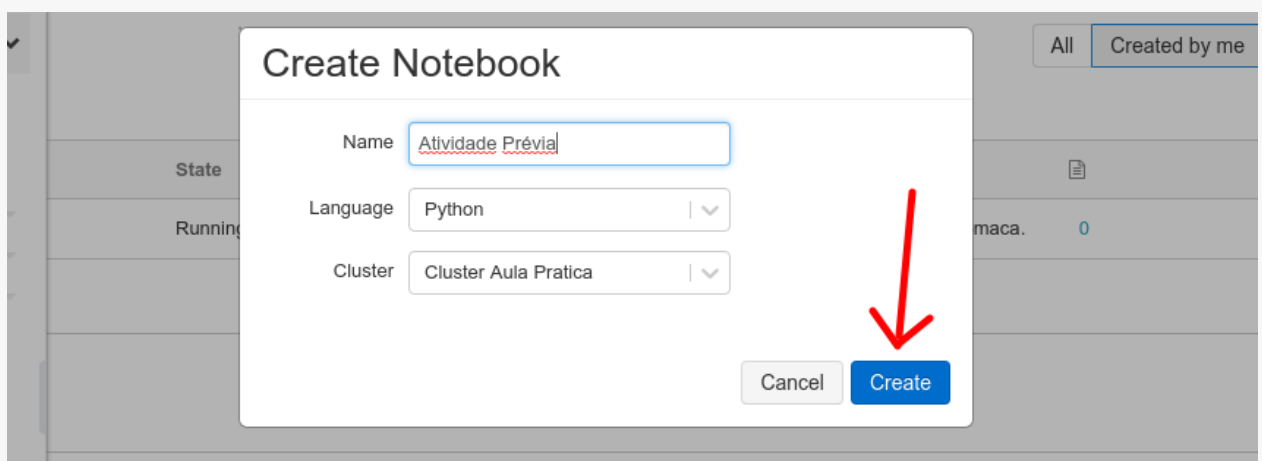
Na próxima aba, adicione um nome para essa pasta.



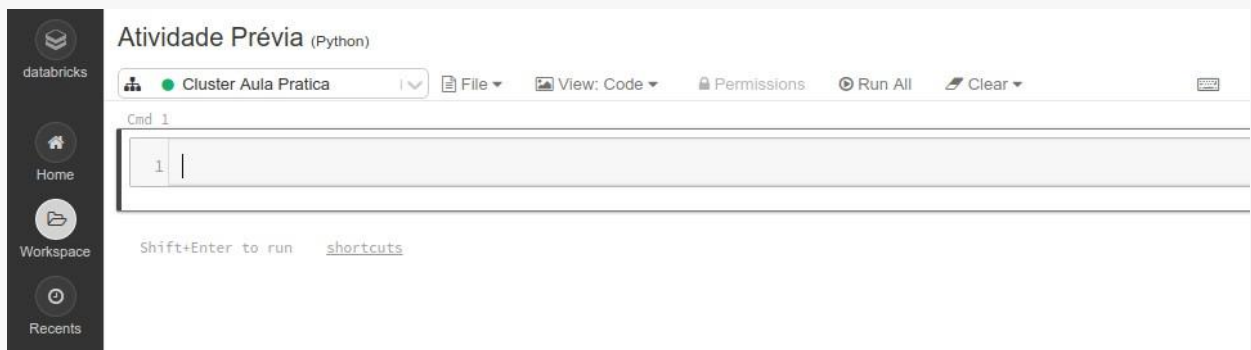
Com a pasta criada, clique na seta em frente ao nome dessa pasta. Selecione “Create” e depois “Notebook”. A figura abaixo apresenta a sequência de passos descrita.



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.

