Hadoop

O Hadoopé uma das ferramentas mais populares quando se trata de processamento de BigData.

Ele é baseado em uma estrutura distribuída que permite o armazenamento e processamento simultâneo de grandesvolumes de dados em clusters de computadores. Além disso, o Hadoop oferece uma biblioteca de análise de dados e uma plataforma para a execução de algoritmos de machinelearning.

****2. ApacheSpark

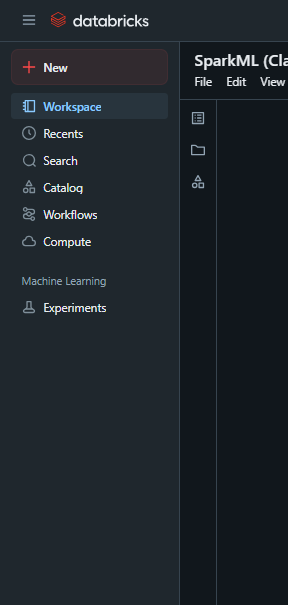
O ApacheSpark é outra ferramenta amplamente utilizada para o processamento de Big Data. Ele oferece um ambiente de análise rápidoeescalável, suportando diferentes linguagens de programação, como Python, Scala e Java.

Além disso, o Spark fornecebibliotecasespecializadas para tarefas como processamento de fluxos de dados em temporeal, análise de gráficos e processamento de machinelearning.

PySpark

PySparkcomDatabricks - Databricks é um site onde podemos compilar nosso código com PySpark sem precisar baixar e instalar localmente, por isso vamos utiliza-lo.

Para isso, acesse o site <https://community.cloud.databricks.com/> e faça o registro do usuário.



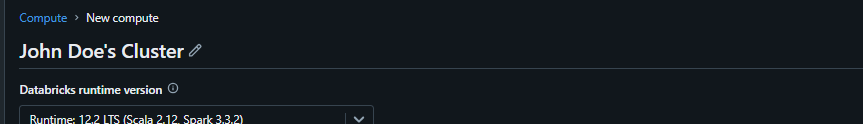
Ao lado temos nossa Sidebar, onde podemos navegar entre as principai ferramentas.

Vamos iniciar com a criação de um novo “Cluster”, que é o termo que define uma arquitetura de sistema capaz combinar vários computadores que irão trabalhar em conjunto para o processamento dos dados.

Na SideBar acesse “Compute”, e logo ao lado em “Create compute”



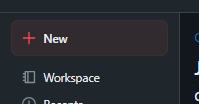
Você pode renomear seu Cluster:



E então cria-lo:

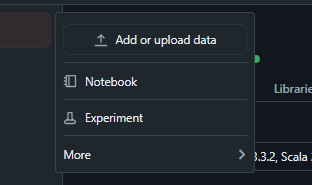


Agora com um Cluster criado podemos criar notebooks onde a análise irá ser definida.



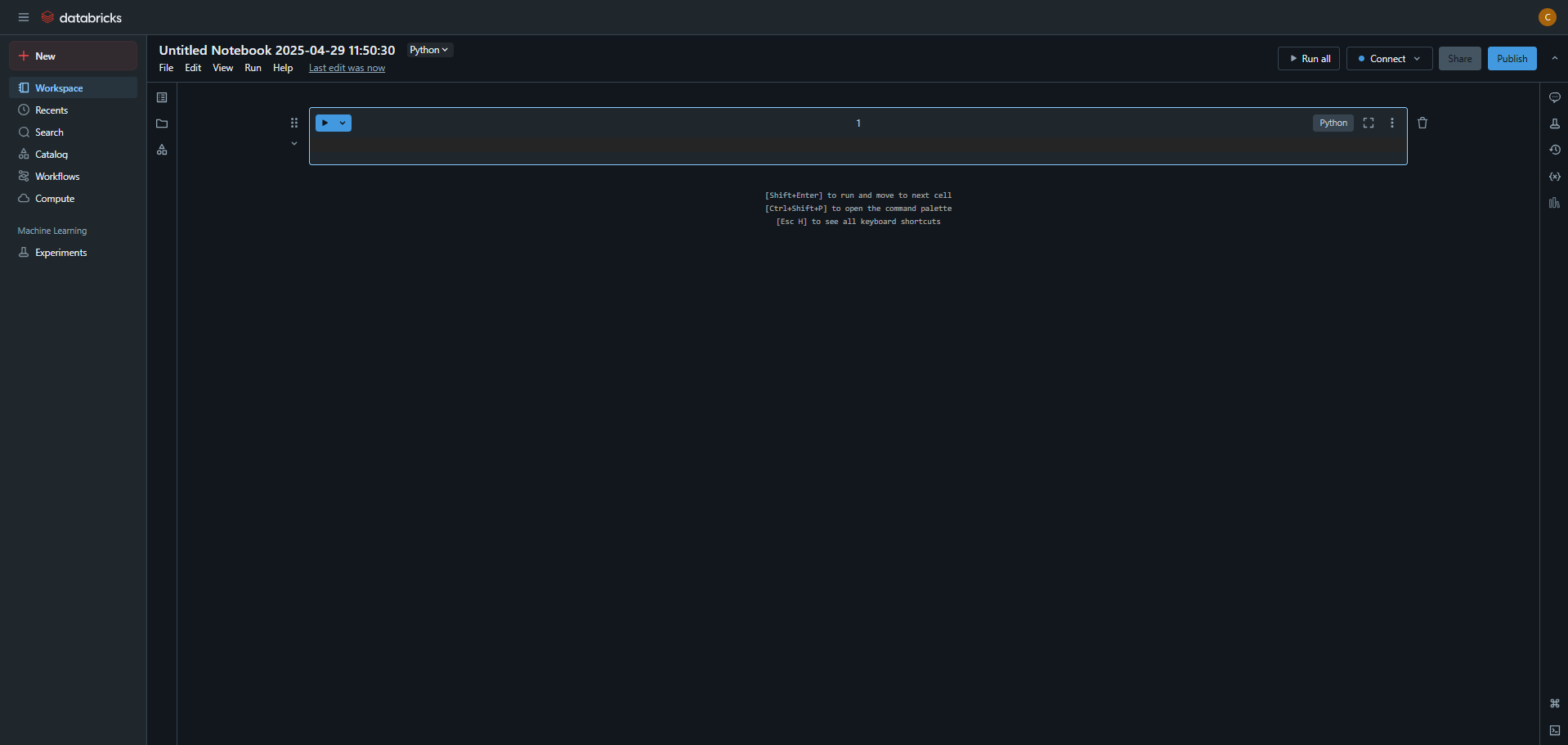
Para criar um notebook, na SideBar aperte

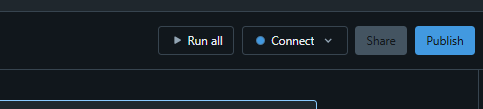
em “+ New”



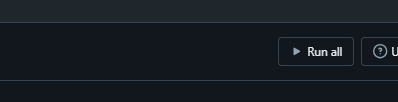
Agora em “Notebook”

Você irá se deparar com a seguinte tela:

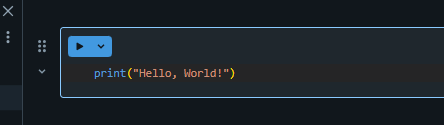


Conecte-se ao Cluster que criou anteriormente clicando em “Connect” 

Para rodar o código pode-se utilizar o atalho “Shift + Enter” para rodar a célula que está focada.



Para rodar todas as células

 Ou então neste botão azul para rodar uma única célula.

Utilizando Dataframes

df = spark.createDataFrame([

    {'Name': 'Alisson', 'age': 28},

    {'Name': 'Donathan', 'age': 22}

])

Para exibir esse dataframe podemos utilizar do método “.show()”:

df.show()

Você pode passar um parâmetro para definir quantas linhas do topo devem ser exibidas:

df.show(8)

Temos também a função “display()” que exibe a tabela de forma mais formatada:

display(df)

Ou então utilizar ambos para aproveitar suas funcionalidades:

display(df.show(8))

Para conseguir o total de linhas no dataframe:

df.count()