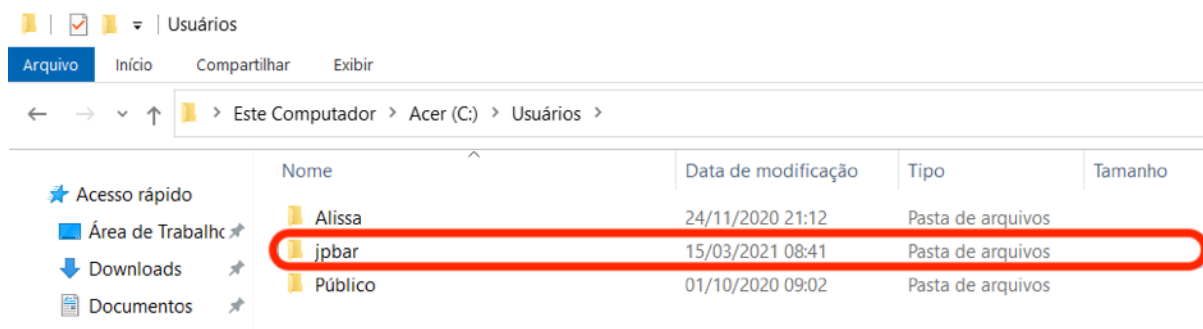


Instruções de Configuração – MapReduce e Spark no Windows

O Hadoop é uma tecnologia que pode ser executada em diferentes sistemas operacionais. Contudo, sua instalação no Windows requer etapas extra que não são demandadas em outros sistemas operacionais. Apesar de simples, qualquer detalhe de configuração mal realizado inviabilizará a implementação e execução de rotinas MapReduce. Ao ler esse tutorial, realize as etapas em seu computador seguindo, **sem exceções** às tratativas sugeridas. A menor divergência nos fará dedicar tempo desnecessariamente a um processo de configuração que é simples.

Etapa 1: Verificação da pasta do usuário

- Abra o “Meu Computador” e acesse seu drive “C:/Usuários/” e encontre o seu usuário.
- Verifique se a pasta do seu usuário possui espaço.



- No exemplo acima, note que o usuário “jpbar” **não** possui nenhum espaço, e portanto, **funcionará** sem problemas.

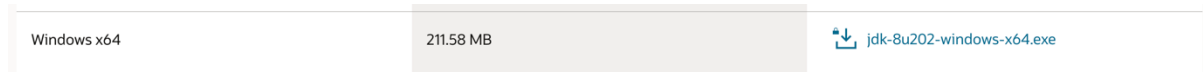
Importante: Caso seu usuário e o nome da pasta **possuam espaço(s)**, você **deverá** criar um usuário em seu computador sem espaços.

Etapa 2: Instalação do Java 8

- Acesse o link a seguir e realize o download do Java JDK 8 de acordo com a arquitetura do seu computador.

Link: <https://www.oracle.com/br/java/technologies/javase/javase8-archive-downloads.html>

- Atualmente, boa parte dos computadores e sistemas operacionais são 64 bits, portanto, prefira a versão “Windows x64”, apresentada abaixo.



Após o download, realize a instalação do Java.

Importante: conforme descrito acima, é necessário termos o Java 8 rodando. Não é Java 7, 11, 12, 13 ou 14. É Java 8!

Etapa 3: Instalação da biblioteca C++ 2020

- Acesse o link a seguir e realize o download do pacote C++ 2010.

Link: https://aka.ms/vs/17/release/vc_redist.x64.exe

- Após o download, realize a instalação.

Etapa 4: Instalação do IntelliJ IDEA

- Acesse a página do IntelliJ IDEA e realize o download e instalação da versão Community pois não requer licença paga.

Link: <https://www.jetbrains.com/pt-br/idea/>

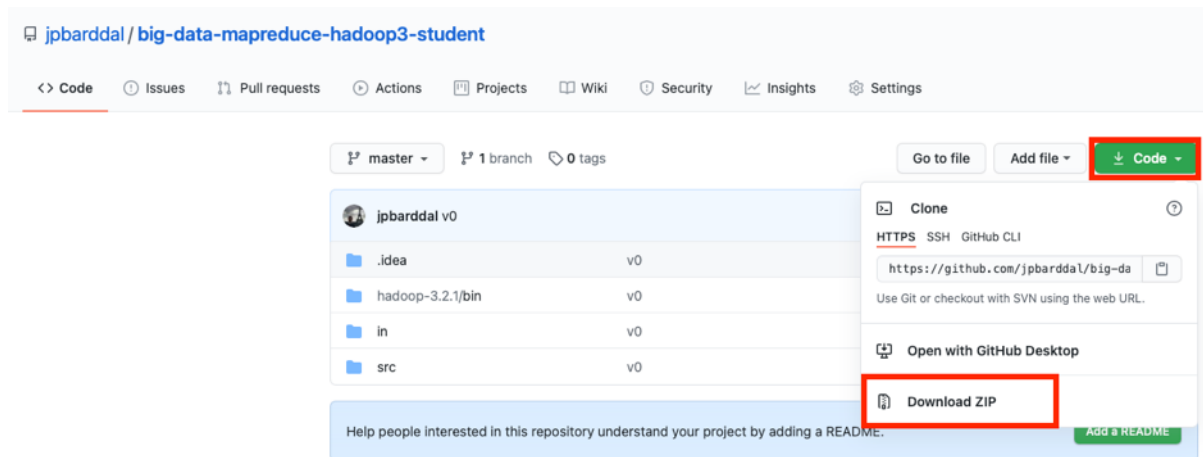
- Após o download, realize a instalação do IntelliJ IDEA.

Etapa 5: Download do projeto de MapReduce

- Realize o download do projeto de MapReduce no link abaixo.

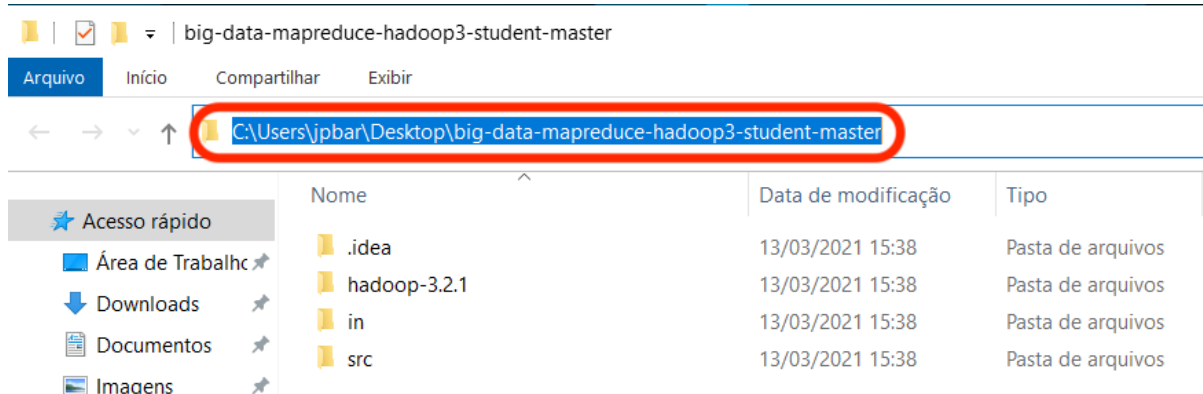
Link: <https://github.com/jpbarddal/big-data-mapreduce-hadoop3-student>

Para realizar o download, sugere-se selecionar a opção “Code > Download ZIP”, conforme representado na imagem abaixo.



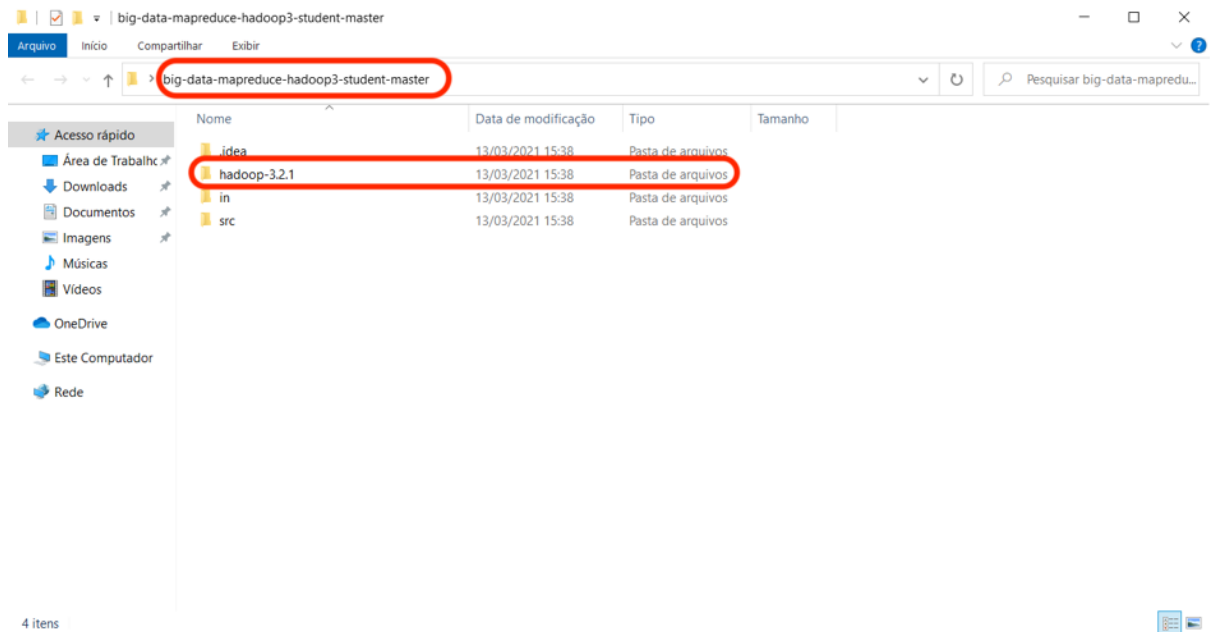
- Extraia a pasta completa em uma pasta de sua preferência.

- Neste tutorial, a pasta foi colocada na “Área de Trabalho” (Desktop), conforme imagem abaixo.

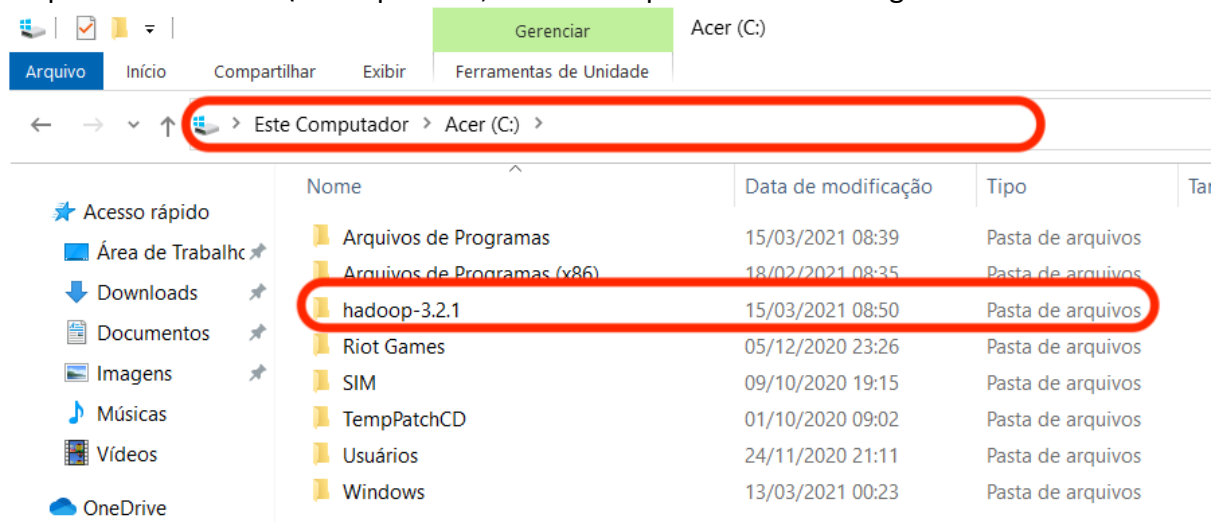


Etapa 6: Configuração do Hadoop e Paths de Sistema

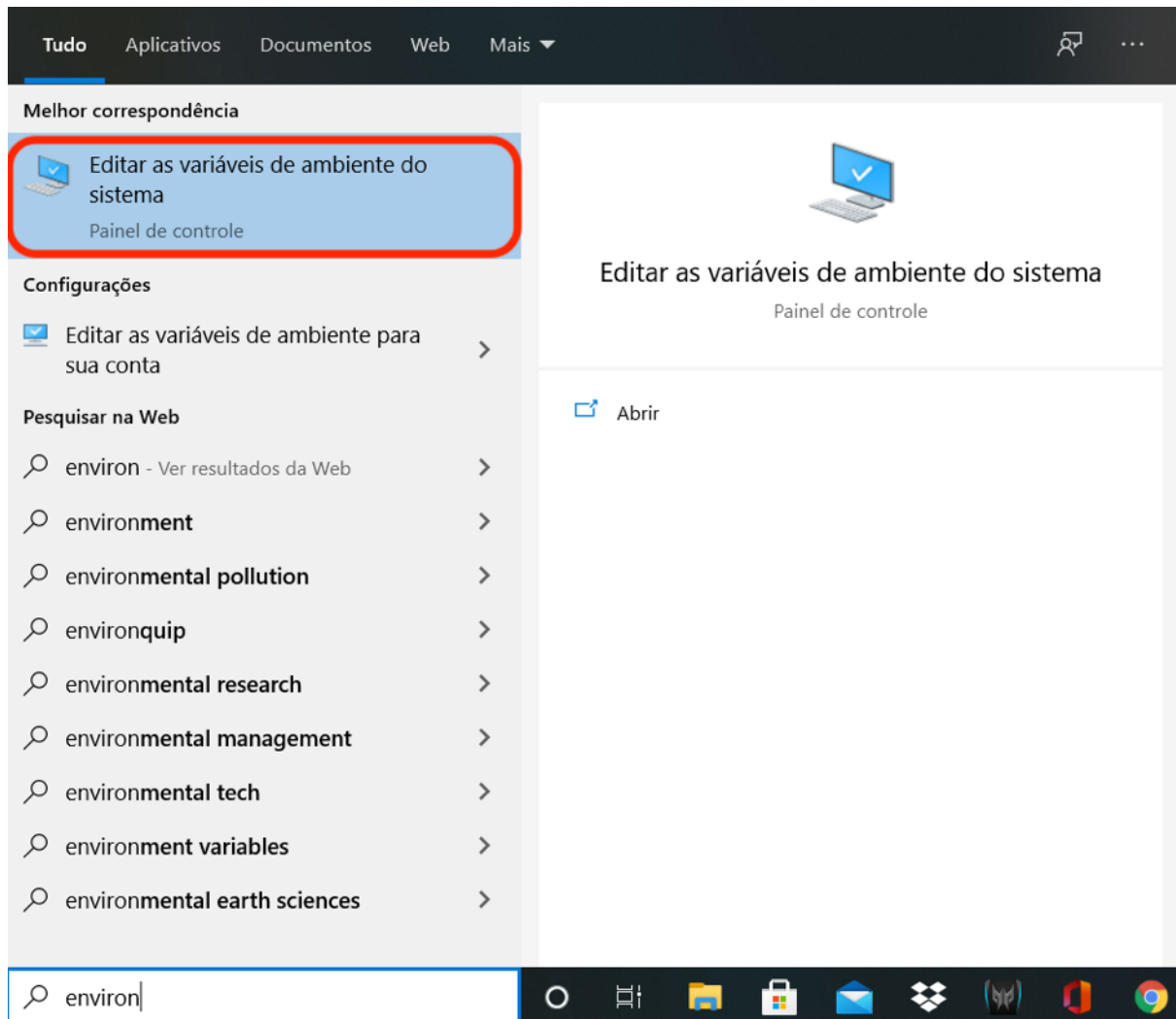
- Abra a pasta do projeto que foi extraída na etapa anterior. Haverá uma pasta com nome no formato “hadoop-X.Y.Z”. No momento da confecção deste tutorial, o nome exato da pasta é “hadoop-3.2.1”, contudo, isso pode mudar ao longo do tempo e conforme o Hadoop e o projeto são atualizados. A imagem abaixo representa a estrutura da pasta e enaltece a pasta “hadoop-3.2.1”.



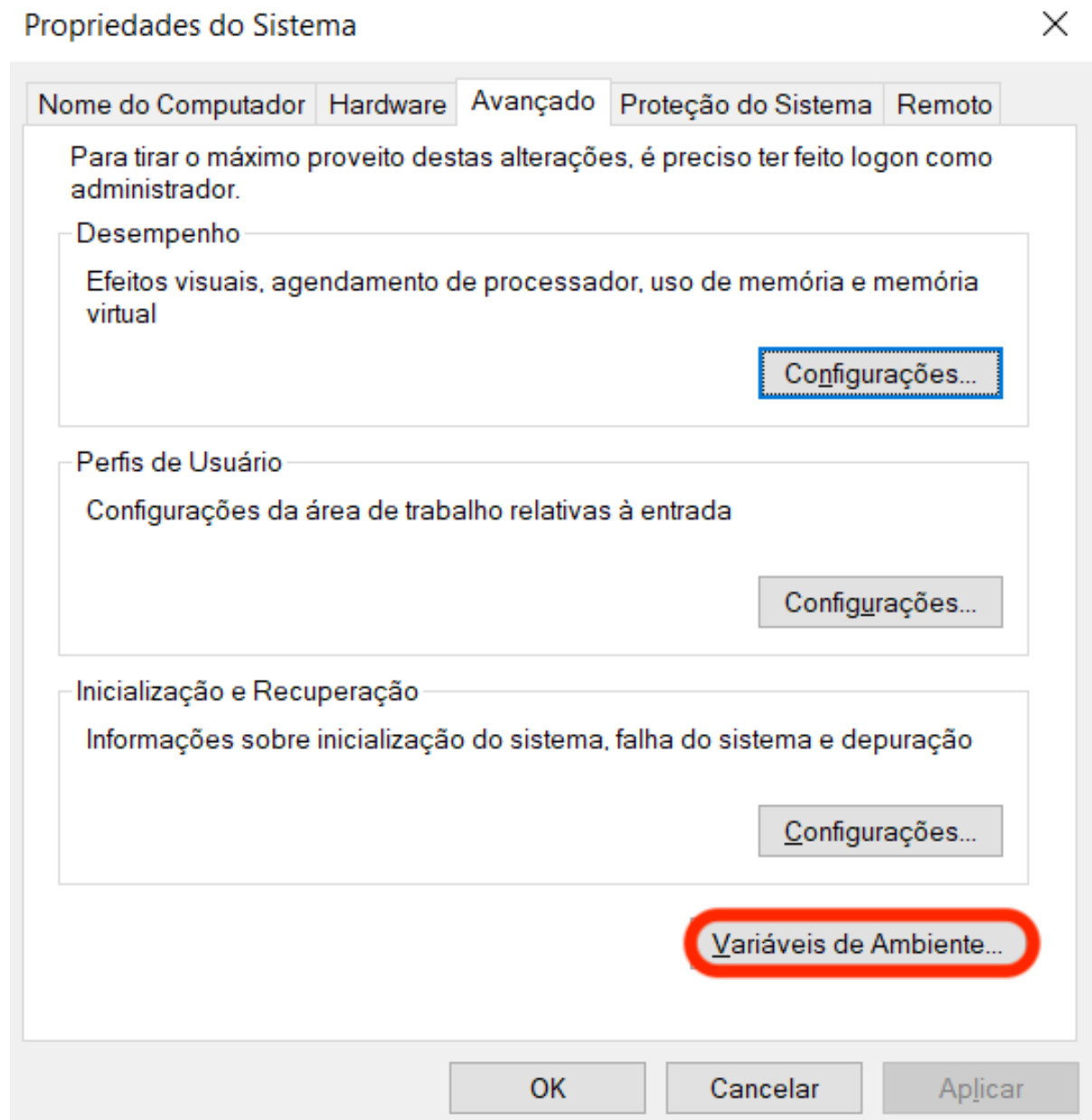
- Copie a pasta “hadoop-3.2.1” para o drive C:\ do seu computador. Desta forma, o caminho da pasta deve ser “C:\hadoop-3.2.1”, conforme apresentado na imagem abaixo:



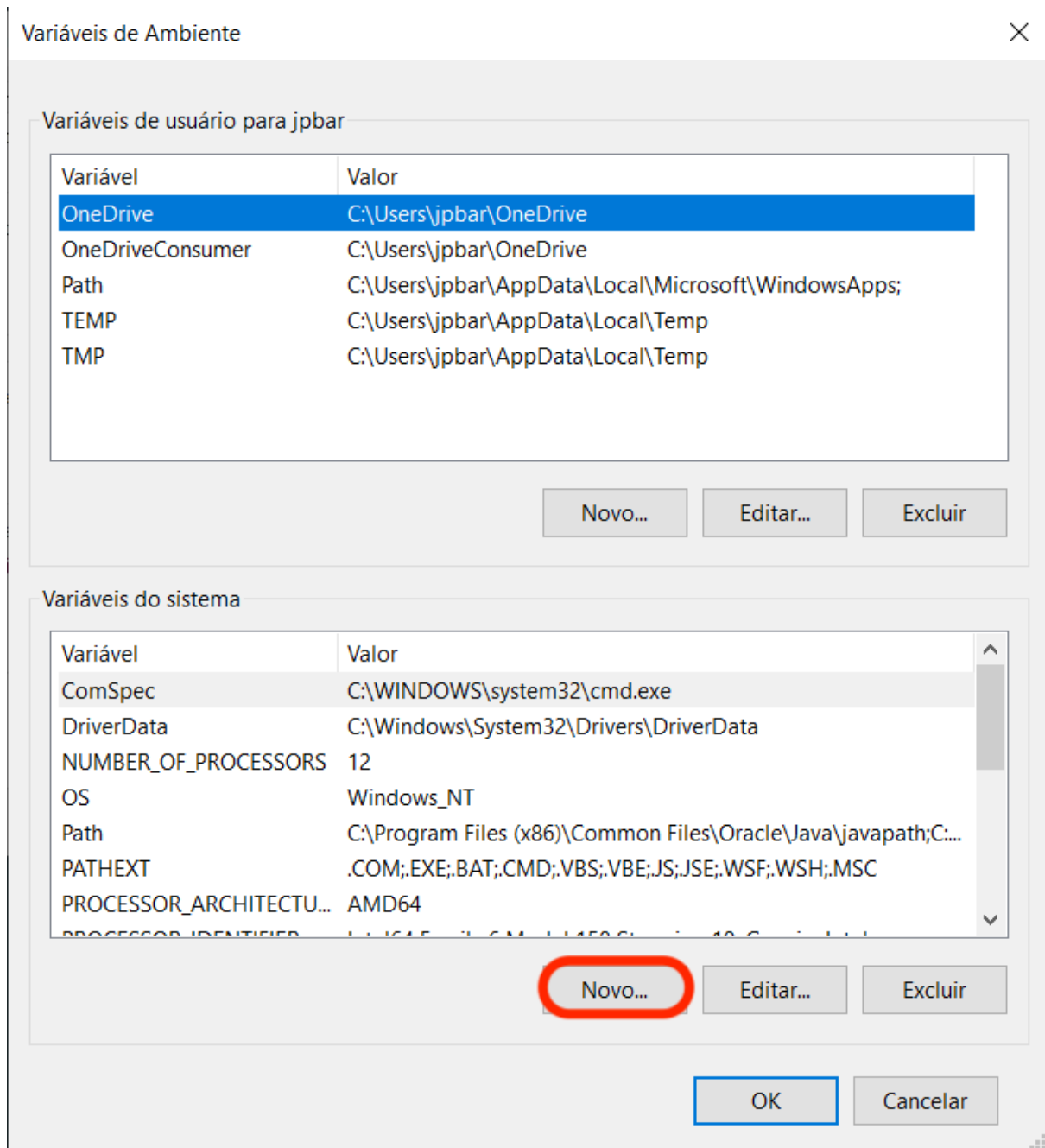
- Abra o menu “Iniciar” e digite “environ”. Você deverá encontrar e clicar na opção apresentada em vermelho na imagem abaixo.



- Uma nova janela abrirá, e você deve clicar em “Variáveis de Ambiente...”.

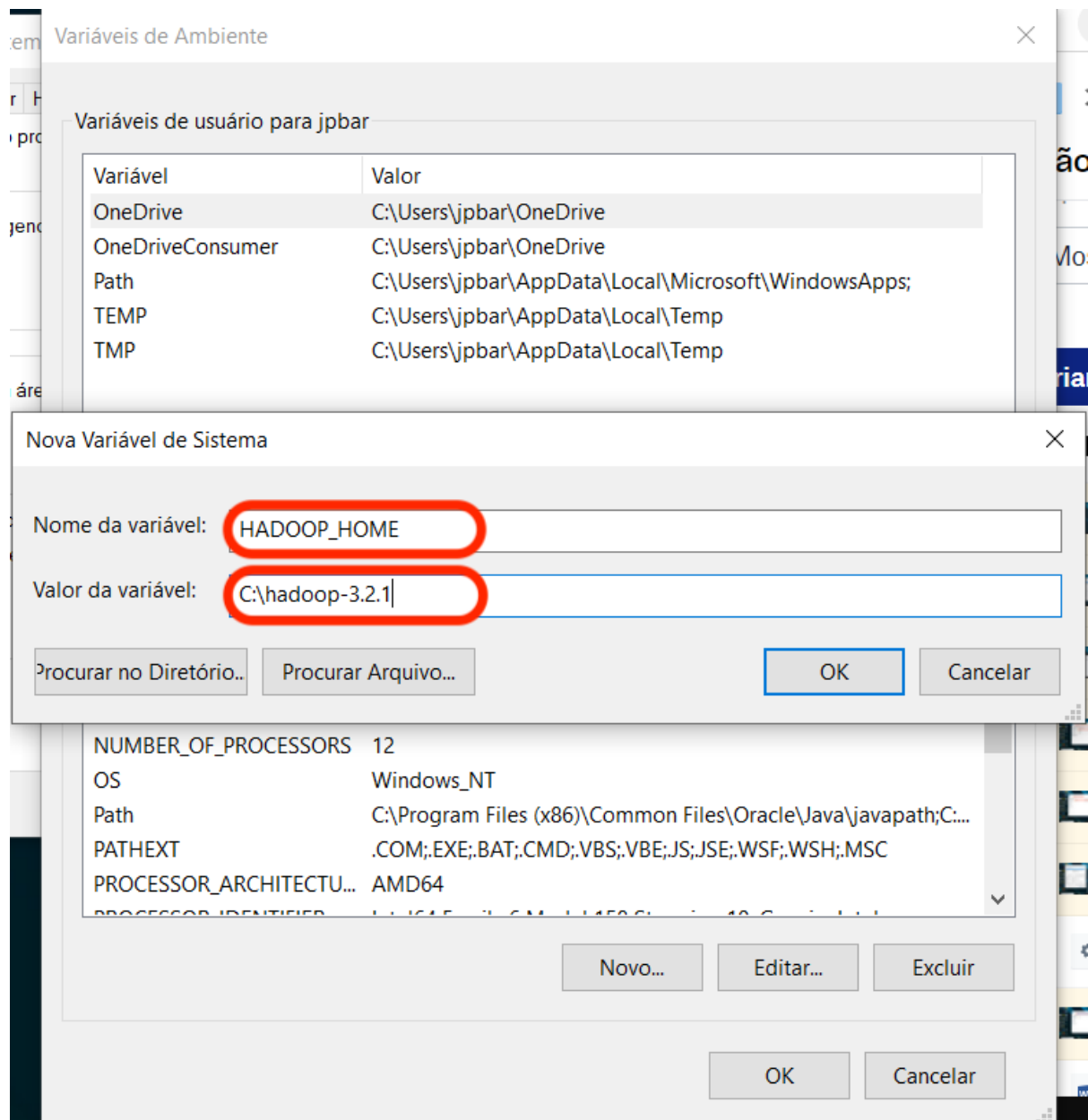


- Posteriormente, na tela descrita na imagem abaixo, clique em “Novo...” na parte de “Variáveis do sistema”.



- Você deve inserir uma nova variável com nome "HADOOP_HOME", e o caminho do diretório é a pasta do Hadoop, definido anteriormente em "C:/hadoop-X.Y.Z". No exemplo, abaixo, o diretório é "C:/hadoop-3.2.1".

Importante: Não é HADDOP, HADOP, HADUP, portanto, verifique se a grafia está correta antes de avançar.



- Na sequência, selecione a variável "Path", também no contexto de "Variáveis do sistema" e clique em "Editar...".

Variáveis de Ambiente



Variáveis de usuário para jpbar

Variável	Valor
OneDrive	C:\Users\jpbar\OneDrive
OneDriveConsumer	C:\Users\jpbar\OneDrive
Path	C:\Users\jpbar\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\jpbar\AppData\Local\Temp
TMP	C:\Users\jpbar\AppData\Local\Temp

Novo... Editar... Excluir

Variáveis do sistema

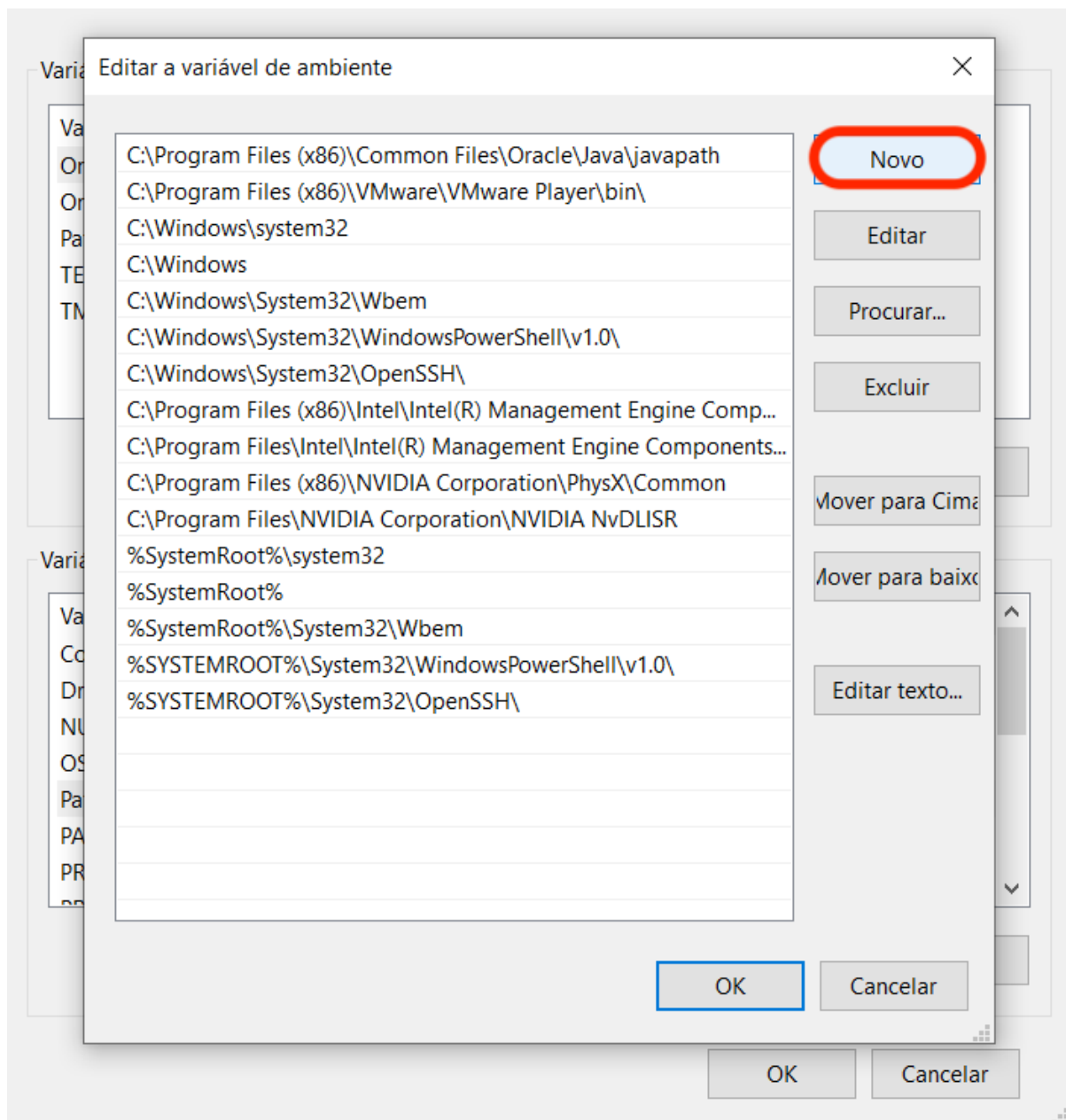
Variável	Valor
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	12
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTU...	AMD64
PROCESSOR_IDENTIFIER	...

Novo... Editar... Excluir

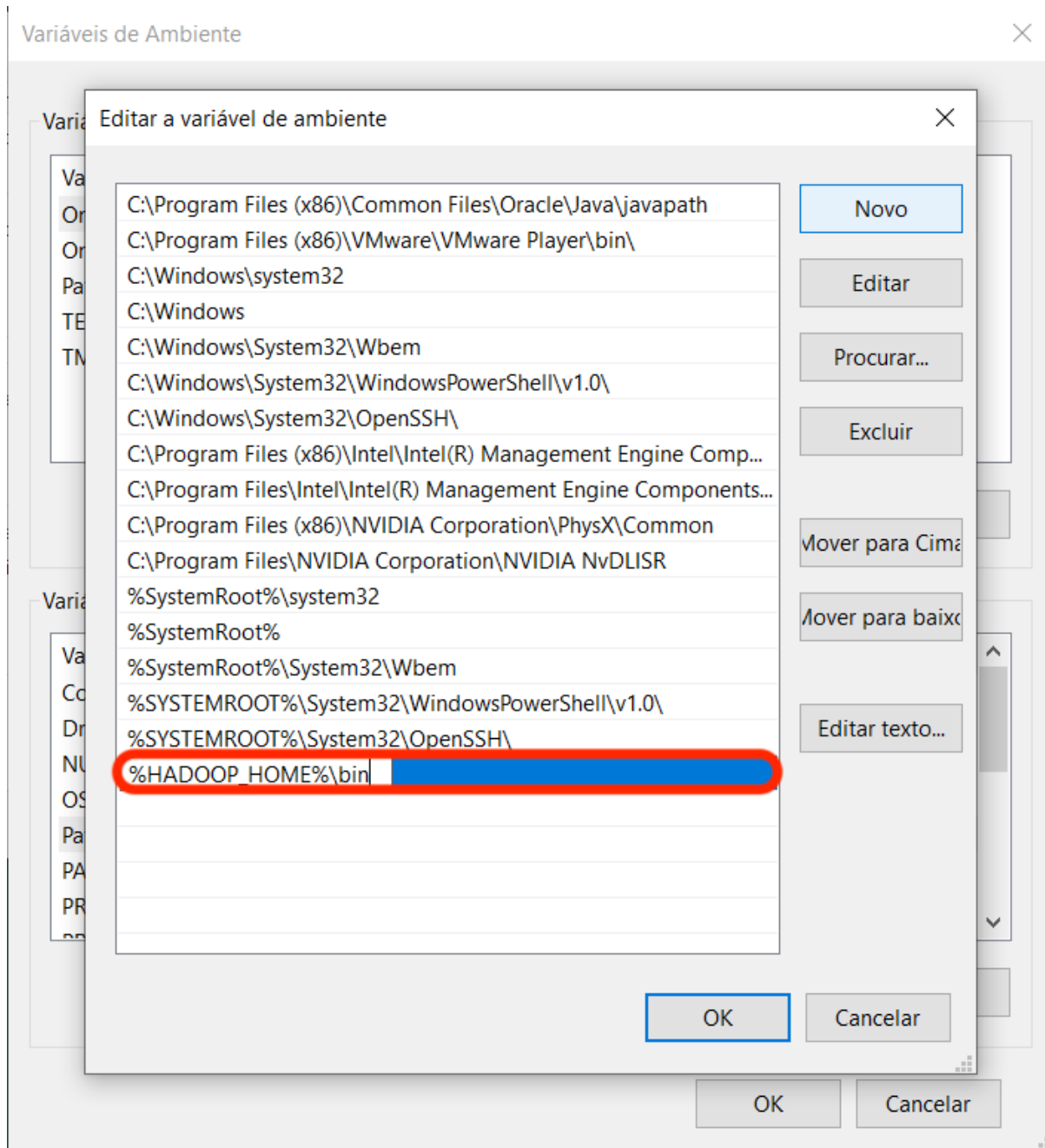
OK Cancelar

- Na janela que abrirá, clique em “Novo”:

Variáveis de Ambiente



- No campo de edição de texto, adicione: “%HADOOP_HOME%\bin”, conforme apresentado na imagem abaixo.



- Dê “OK” em todas as janelas e reinicie o computador.

MUITO IMPORTANTE: Se este tutorial estiver sendo usado para configuração dos computadores da PUCPR, é necessário fornecer permissões de escrita e leitura para a pasta C:\tmp.

Marcelo de Jesus e Leonardo definiram os seguintes comandos para isso:

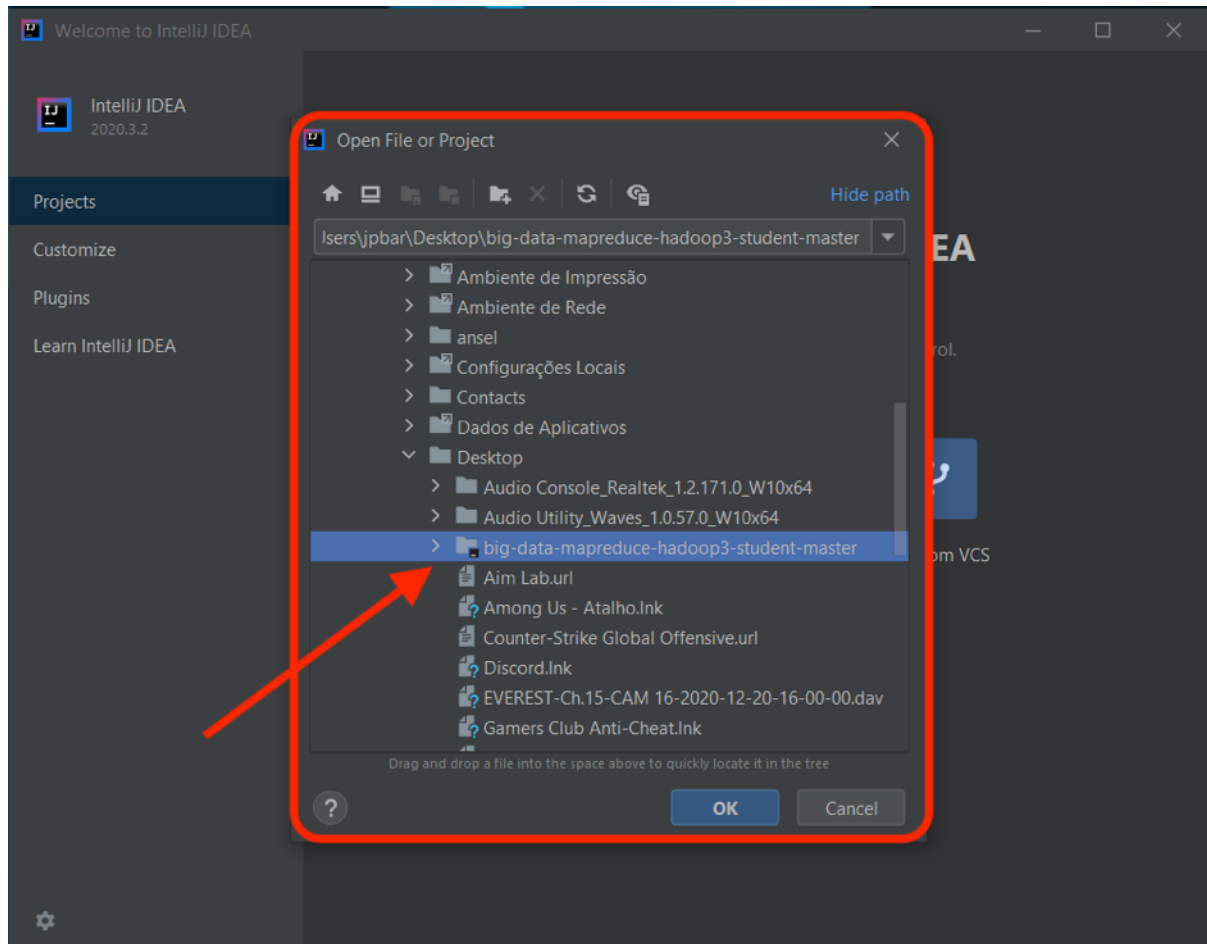
```
cacls "C:\tmp\hadoop\mapred\staging" /E /P Todos:F
```

```
cacls "C:\tmp" /E /P Todos:F
```

--- Caso tenham dúvidas, favor convocar Marcelo de Jesus para esta etapa.

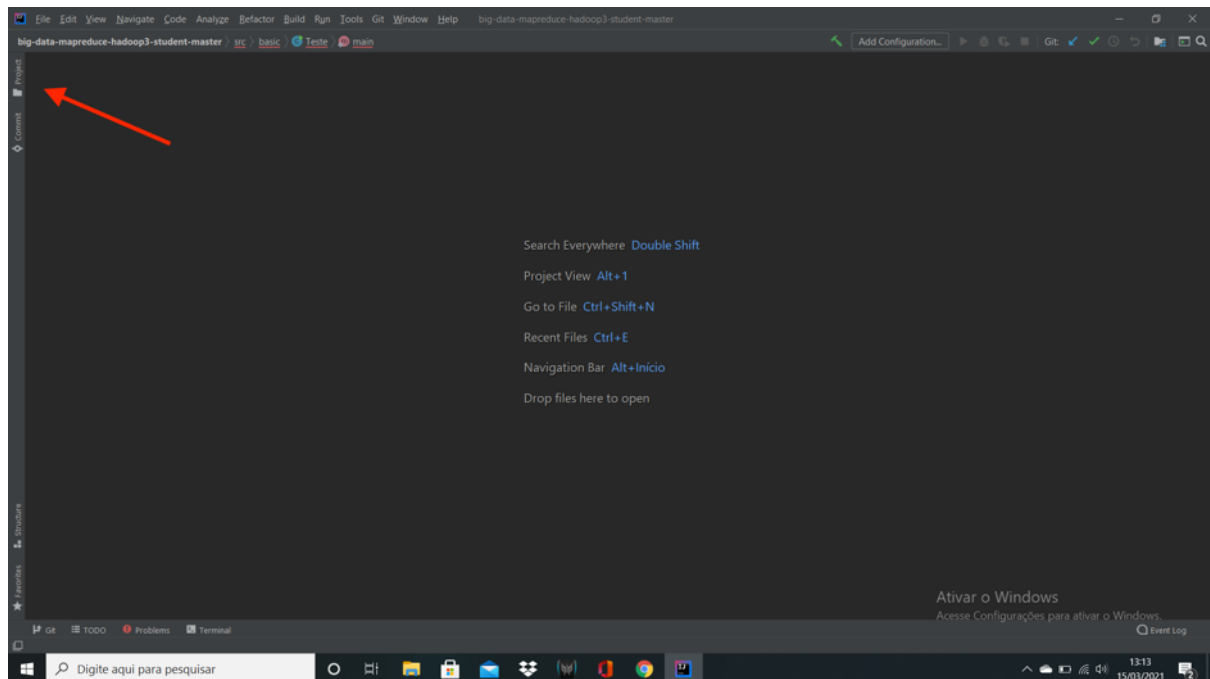
Etapa 7: Teste

- Inicialize o IntelliJ IDEA.
- Na janela abaixo, clique em Open e selecione a pasta do projeto. **Importante:** Note que a pasta do projeto possui um ícone diferenciado, uma vez que há um quadrado preto no canto inferior direito do ícone, diferentemente das demais pastas.

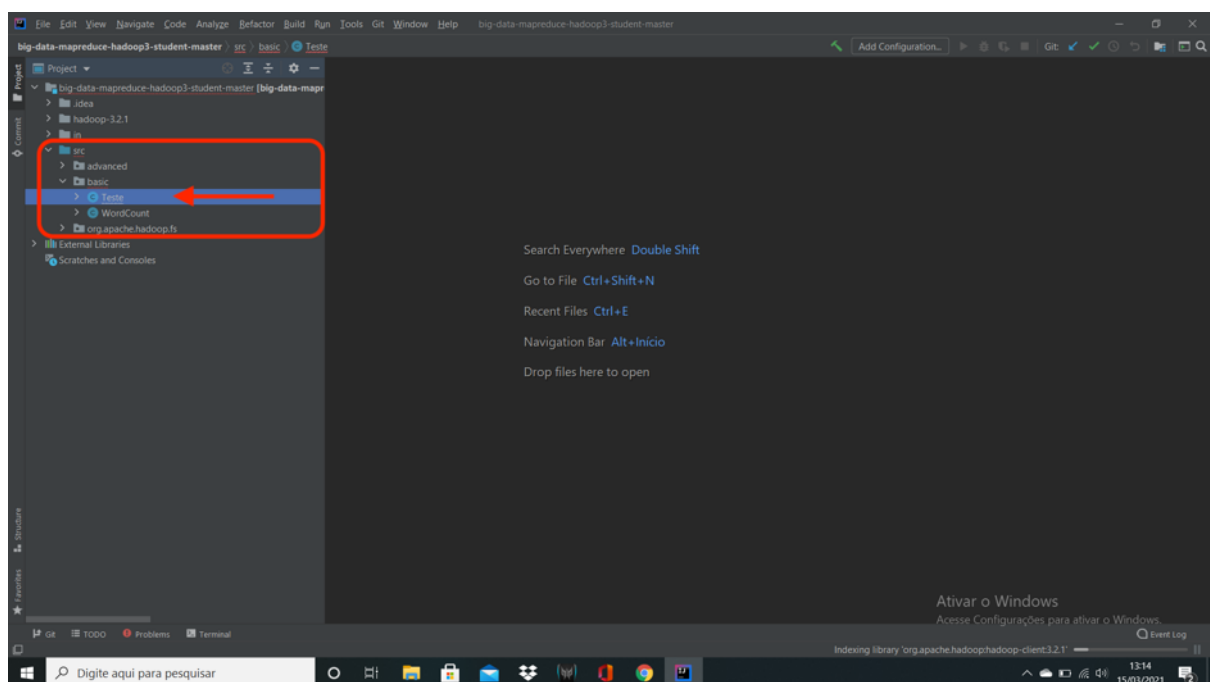


- Aguarde até que a barra de carregamento apresentada no canto inferior direito termine de carregar o Java.

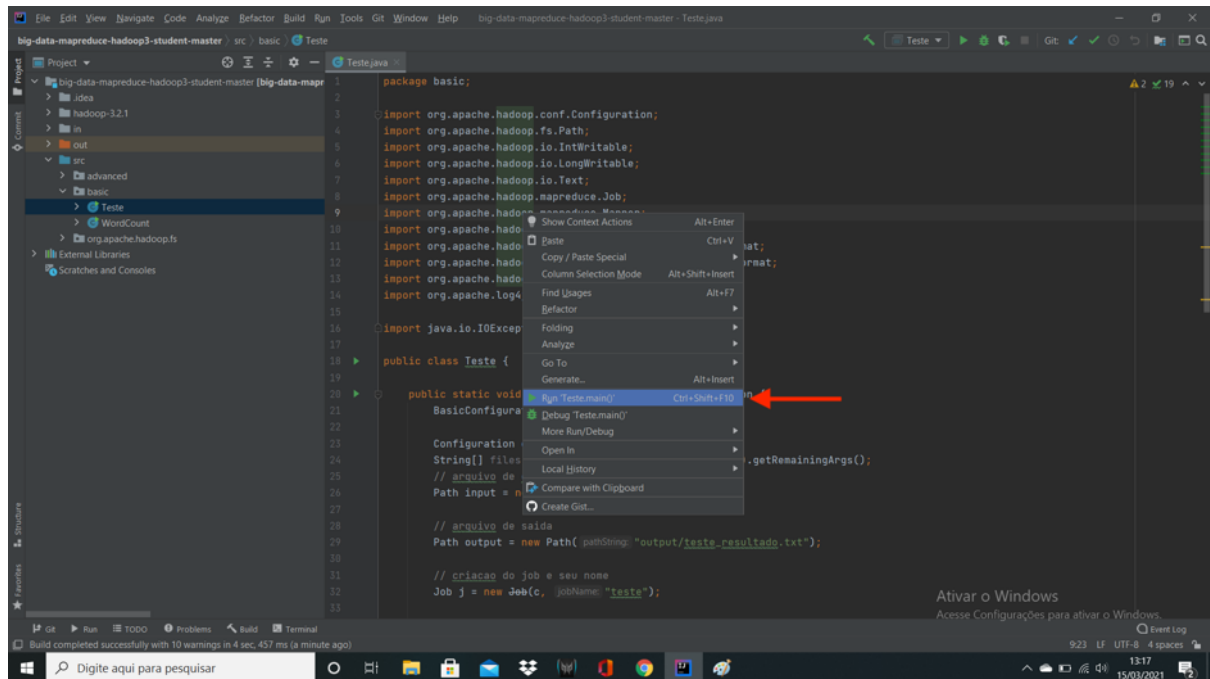
- Abra a aba de Projeto, disponível na parte esquerda da interface.



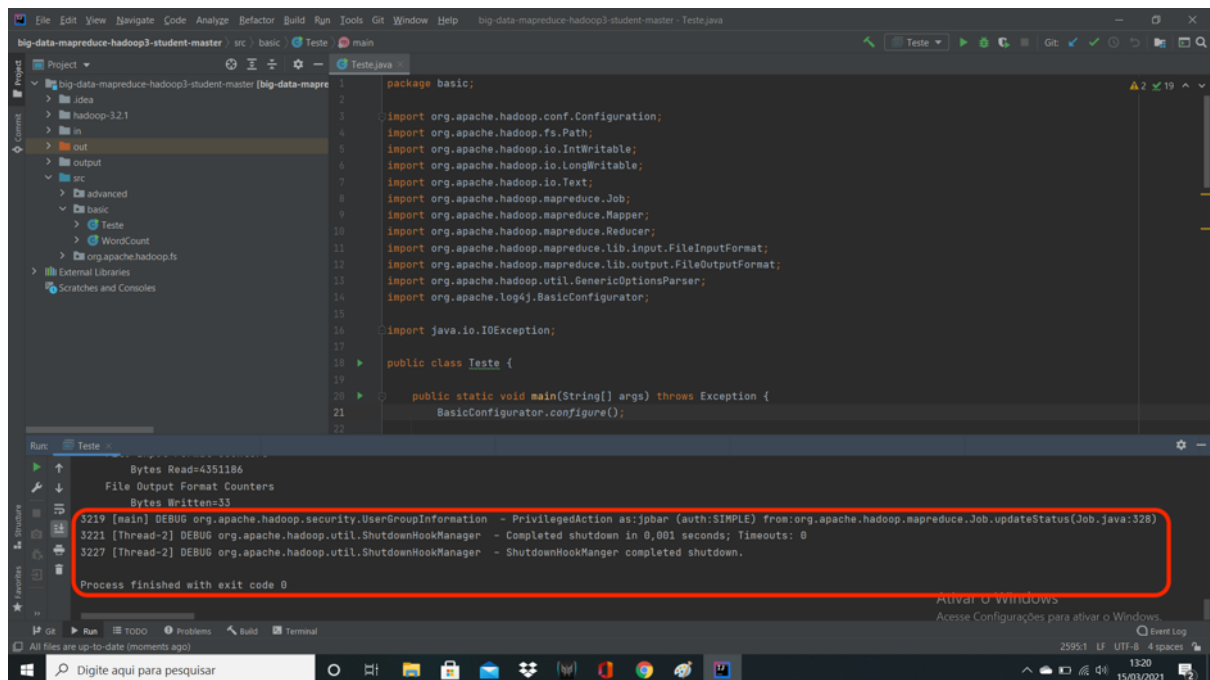
- Abra a classe “Teste.java”, disponível em “src > basic > Teste”.



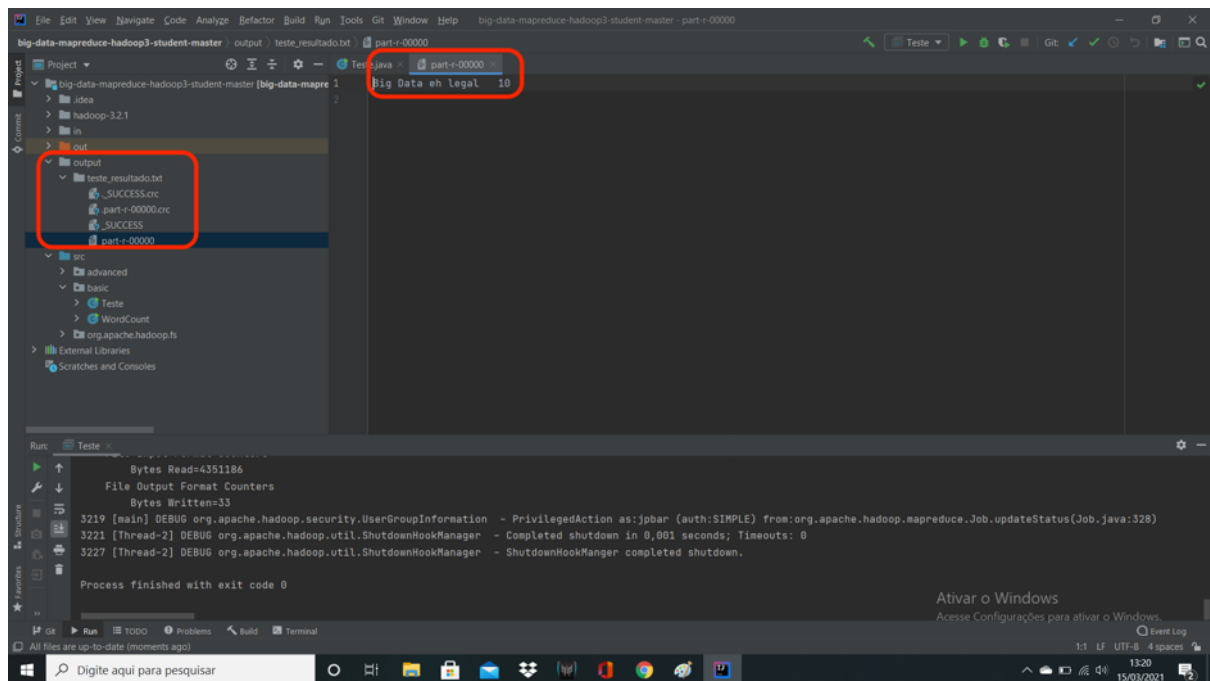
- Com o código-fonte da classe aberta, clique com o botão direito em cima de qualquer parte do código e clique em “Run”. A primeira vez tende a demorar um pouco mais pois o IntelliJ fará o download das bibliotecas necessárias do Hadoop.



- Verifique a saída apresentada no console. No final, você deve encontrar a mensagem apresentada abaixo, terminando com “Process finished with exit code 0” (o que significa que nenhum erro aconteceu):



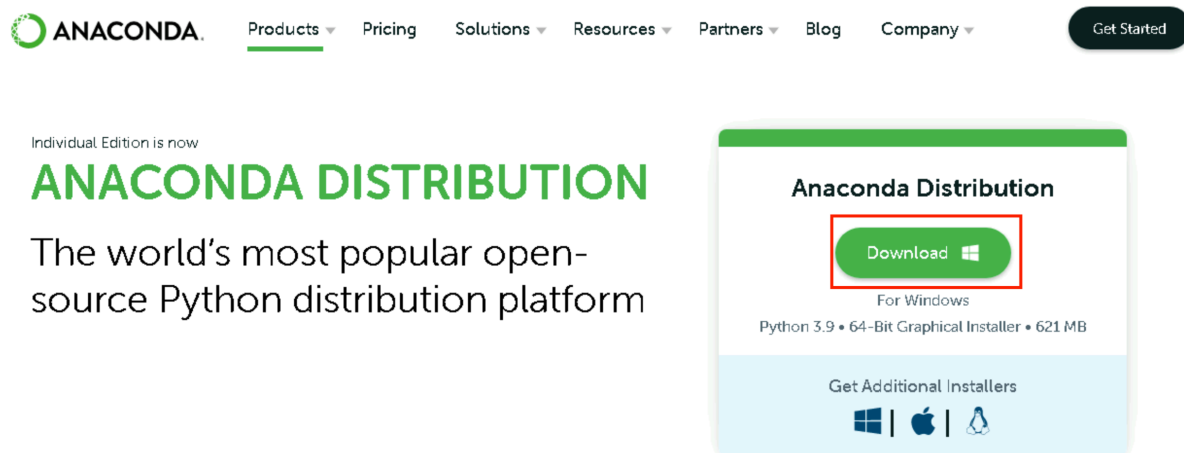
- Finalmente, volte a aba do projeto e abra a pasta “output > teste_resultado.txt > part-r-0000.txt”, conforme descrito abaixo. Esse arquivo deve possuir o conteúdo “Big Data eh legal 10”.



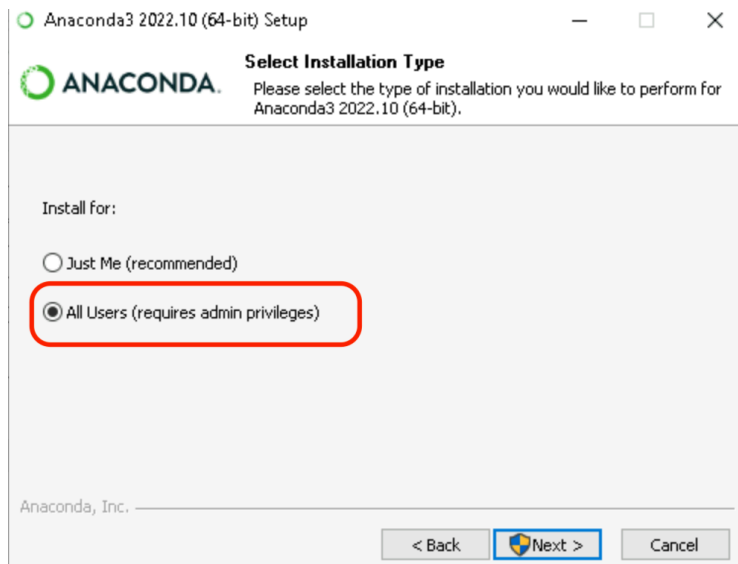
- Se você chegou a este ponto, parabéns, seu computador está configurado para executar MapReduce em modo local. Caso contrário, verifique as etapas anteriores com muita calma. Caso ainda assim você não consiga configurar o ambiente corretamente, verifique os erros comuns apresentados a seguir ou entre em contato com o professor.

Etapa 8: Download e Instalação do Anaconda

- Acesse <https://www.anaconda.com/products/distribution> e faça download do Anaconda para ambiente local.

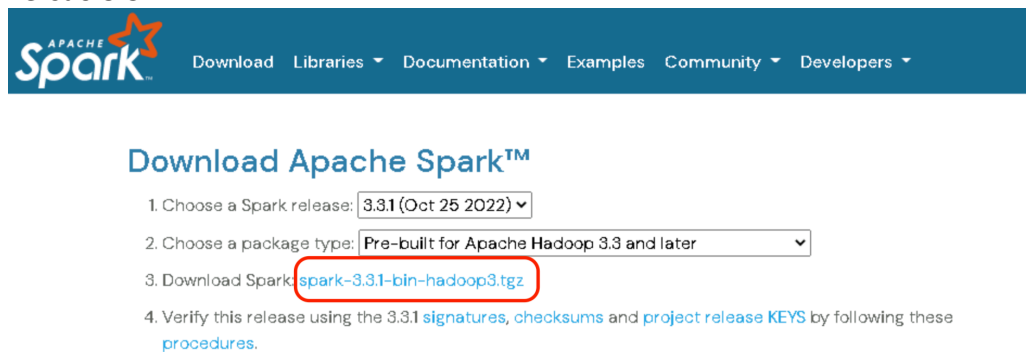


- Durante a instalação, selecione a opção para instalar para todos os usuários “All Users”.



Etapa 10: Download e Configuração do Spark

- Acesse o site do Spark (<https://spark.apache.org/downloads.html>) e faça download da versão 3.3.1:

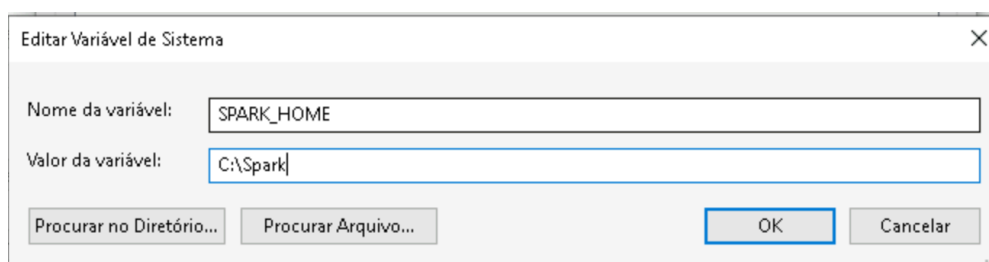


- Extraia o arquivo .tgz resultante do download. Sugestão: use o 7zip para extrair a pasta.

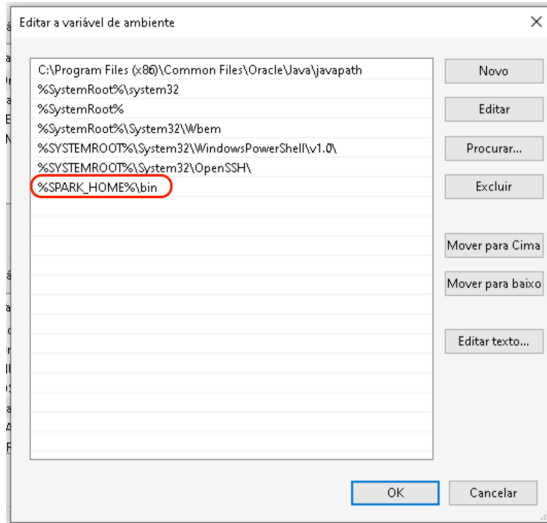
- Mude o nome da pasta extraída para Spark

- Mova a pasta para C:\Spark

- De forma similar ao HADOOP, crie uma variável de ambiente para o sistema chamada SPARK_HOME com o diretório C:\Spark



- Dentro da variável PATH, adicione o caminho %SPARK_HOME%\bin

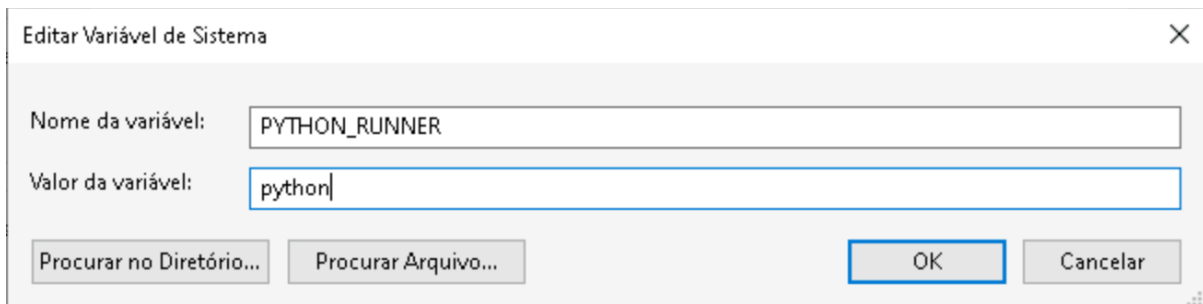


- Ainda dentro da PATH do sistema, adicione o caminho do python:

C:\Users\projeto\.conda\envs\bigdata\

Importante: verifique o path do python relacionado ao ambiente big data gerado no conda.






- Adicione ainda uma variável de ambiente chamada PYTHON_RUNNER com valor igual a "python".



- Nas configurações do computador, abra a opção "Aliases de execução de Apps" e desabilite as componentes relativas ao Python

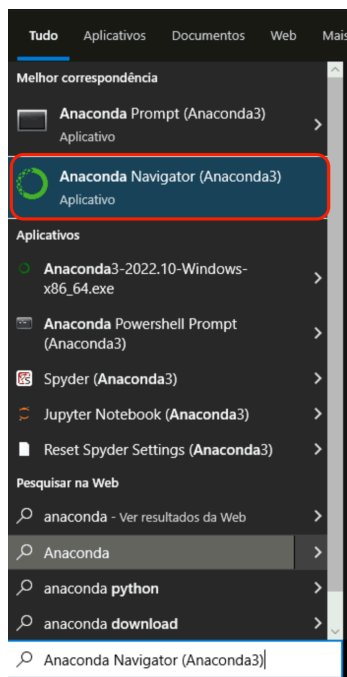
🏠 Aliases de execução de app

Os aplicativos podem declarar um nome usado para executar o aplicativo em um prompt de comando. Se vários aplicativos usarem o mesmo nome, escolha o nome a usar.

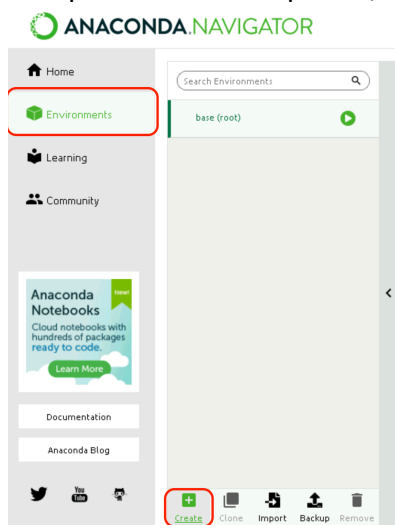
	Skype Skype.exe	<input checked="" type="checkbox"/>	Ativado
	Windows Package Manager Client winget.exe	<input checked="" type="checkbox"/>	Ativado
	Xbox Game Bar GameBarElevatedFT_Alias.exe	<input checked="" type="checkbox"/>	Ativado
	App Installer python.exe	<input type="checkbox"/>	Desativado
	App Installer python3.exe	<input type="checkbox"/>	Desativado

Etapa 9: Criação de Ambiente

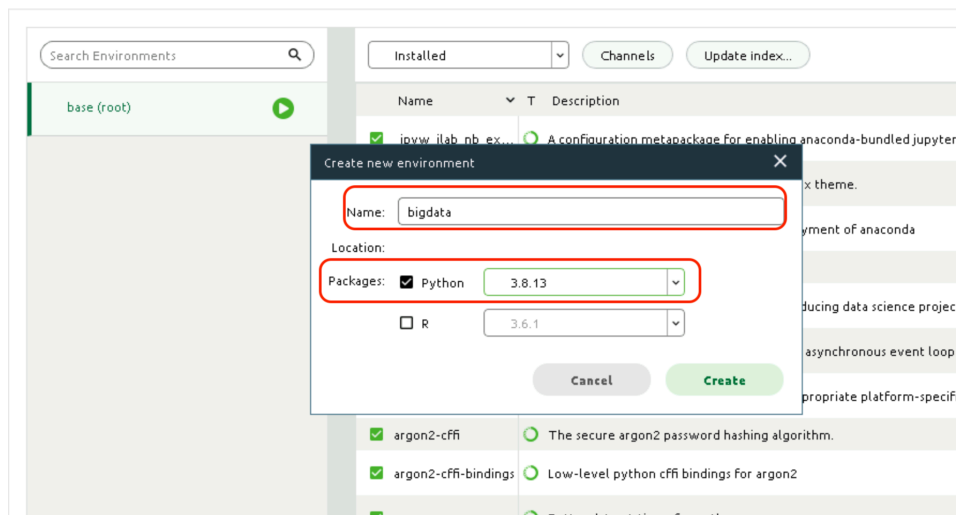
- Inicie a interface do Anaconda chamada "Anaconda Navigator".



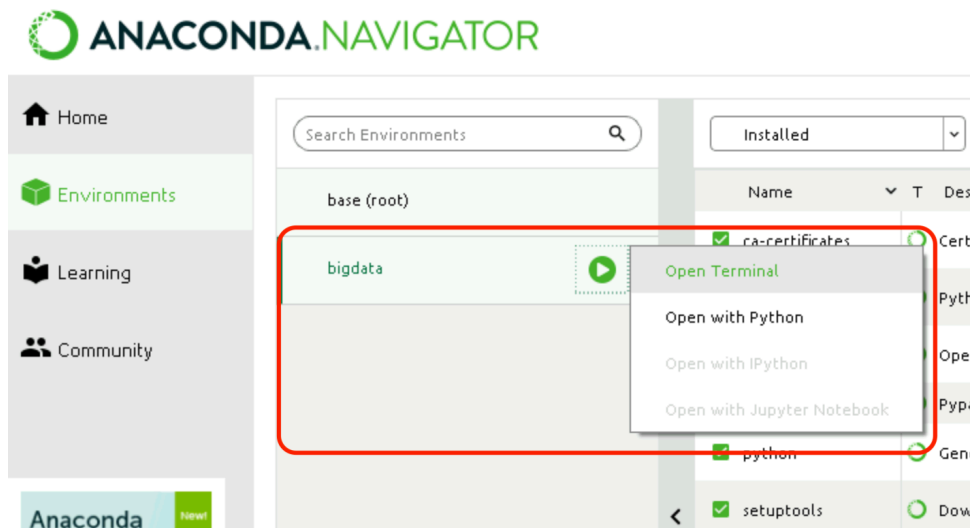
- Do lado esquerdo, clique em “Environments”.
- Na parte inferior esquerda, clique em “Create”.



- Na janela que abrirá, crie um ambiente com o nome “bigdata” e versão do Python 3.8.



- Após a criação, clique no botão verde ao nome do ambiente e depois em “Open Terminal”.

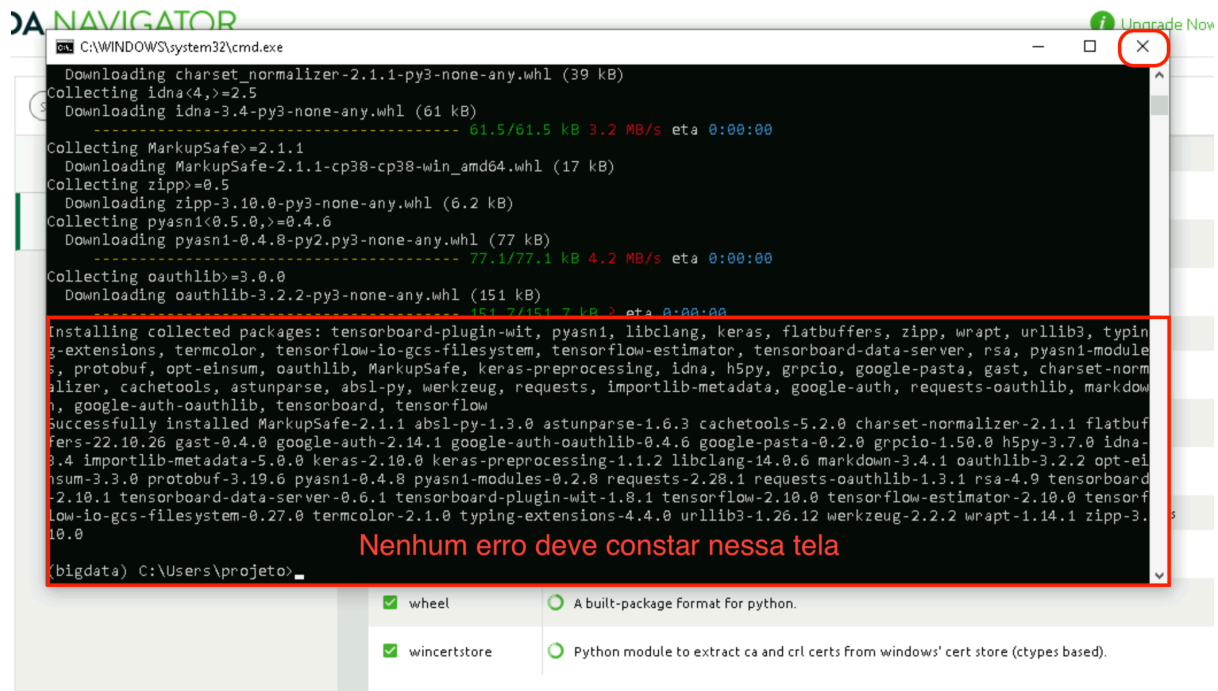


- Na tela do terminal aberta, digite, sem ressalvas:

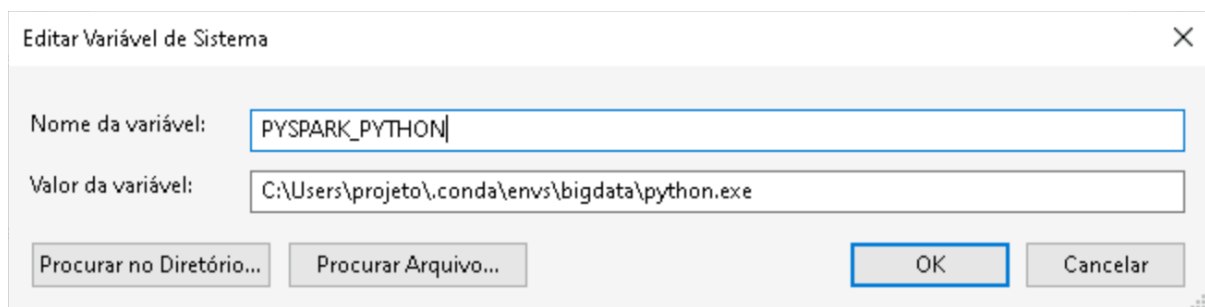
`“pip install pandas imblearn scikit-learn numpy scipy matplotlib seaborn pyarrow pyspark==3.3 tensorflow keras”`

- Aguarde a instalação.

- Com a instalação terminada, feche a tela do terminal.

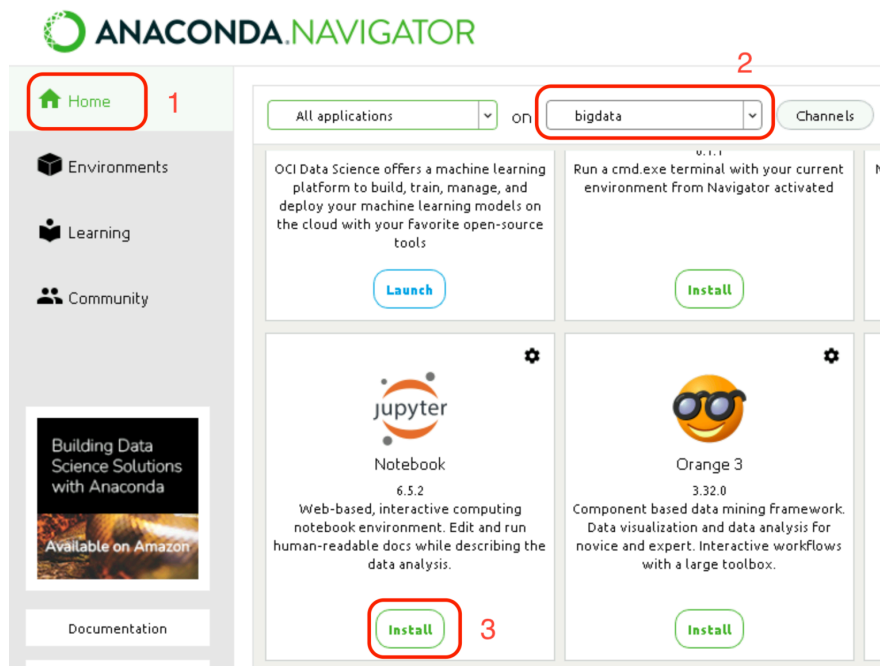


- Ainda com a janela do terminal aberta, execute o comando “where python”. Este comando retornará o path em que o Python está instalado. Adicione uma variável de ambiente a nível de sistema com nome PYSARK_PYTHON e valor igual ao path do ambiente “bigdata”, ligando o arquivo de execução do Python:



-Feche o terminal.

- Do lado esquerdo da interface do Anaconda, clique em “Home”.
- Garanta que o ambiente selecionado seja “bigdata”.
- Encontre o ícone do Jupyter Notebook e clique em Install.



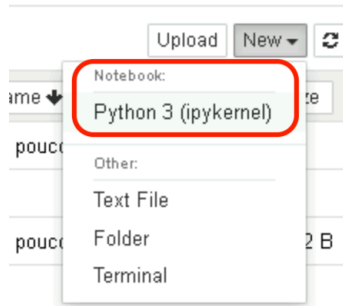
- Aguarde a instalação.

Etapa 10: Teste do PySpark

- Ainda na tela de Home, clique em “Launch” no ícone do Jupyter notebook.



- Isso abrirá uma tela no navegador, conforme apresentado abaixo. Nesta tela, navegue até um diretório com permissão de escrita e crie um notebook, usando o botão “New > Python 3”



- Na tela aberta, digite o seguinte código na primeira linha e use SHIFT+ENTER para executar.

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.master('local').getOrCreate()
import pyspark.pandas as os
spark.sparkContext.parallelize([1,2,3,4]).reduce(lambda x,y: x+y)
```

- Logo abaixo do campo digitado, deverá aparecer o resultado “10”. Abaixo um exemplo da execução:

```
In [2]: from pyspark.sql import SparkSession
spark = SparkSession.builder.master('local').getOrCreate()
import pyspark.pandas as os
spark.sparkContext.parallelize([1,2,3,4]).reduce(lambda x,y: x+y)

WARNING:root:'PYARROW_IGNORE_TIMEZONE' environment variable was not set. It is required to set this environment variable to
'1' in both driver and executor sides if you use pyarrow>=2.0.0. pandas-on-Spark will set it for you but it does not work i
f there is a Spark context already launched.

Out[2]: 10
```