

## 16.4.1

## PySpark in Google Colab Notebooks

**Your** client requests full transparency so that they can eventually maintain this project on their own in the future. Therefore, you'll need to perform your work in a place that is accessible beyond your local laptop. The solution to this will be using cloud-based notebooks.

Before we can begin using Spark, we need a place to do so. **Cloud-based notebooks** provide a remote workspace with stronger resources than our local laptop might allow. Cloud notebooks permit us to share our work with others, such as coworkers, similar to GitHub.

We'll use [Google Colaboratory](https://colab.research.google.com/notebooks/welcome.ipynb) (<https://colab.research.google.com/notebooks/welcome.ipynb>) (Colab), which are Google-hosted notebooks. Some setup is required, so we'll start there before getting back to the basics of PySpark.


First, you'll need a Google account. If you don't already have an account, be sure to sign up for one.


Once you have created a Google account and signed in, navigate to the [Google Colaboratory](https://colab.research.google.com/) (<https://colab.research.google.com/>) webpage. Once on the page you will see a menu like the one below.


[Examples](#)[Recent](#)[Google Drive](#)[GitHub](#)[Upload](#)


Filter notebooks


Title


 Overview of Colaboratory Features


 Markdown Guide


 Charts in Colaboratory


 External data: Drive, Sheets, and Cloud Storage


 Getting started with BigQuery





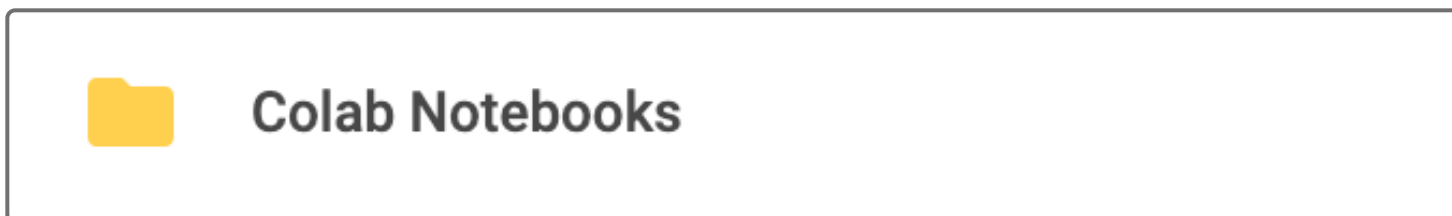






[NEW NOTEBOOK](#)[CANCEL](#)

Click **NEW NOTEBOOK** which will then redirect to you a fresh new notebook to use. Google will store all the notebooks you create in a folder called **Colab Notebooks** on your Google Drive



A new tab will launch with a new notebook. The functionality is very similar to using Jupyter Notebook, except now everything is hosted online.

Just like we do with our local environment, we need to install packages for libraries we want to use. PySpark does not come native to Google Colab and needs to be installed. We'll do so by running the following code in the first cell:

```
import os
# Find the latest version of spark 3.0 from http://www.apache.org/dist/spark/ and enter as the spark version
# For example:
# spark_version = 'spark-3.0.3'
spark_version = 'spark-3.<enter version>'
os.environ['SPARK_VERSION']=spark_version

# Install Spark and Java
!apt-get update
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q http://www.apache.org/dist/spark/$SPARK_VERSION/$SPARK_VERSION-bin-hadoop2.7.tgz
!tar xf $SPARK_VERSION-bin-hadoop2.7.tgz
!pip install -q findspark

# Set Environment Variables
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/{spark_version}-bin-hadoop2.7"

# Start a SparkSession
import findspark
findspark.init()
```

## NOTE

Don't be scared by this installation code. All this does is install Spark for our notebook to use. It can be copied and pasted into any new notebooks you wish to run.

We are now up and running with cloud notebooks with PySpark.

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.