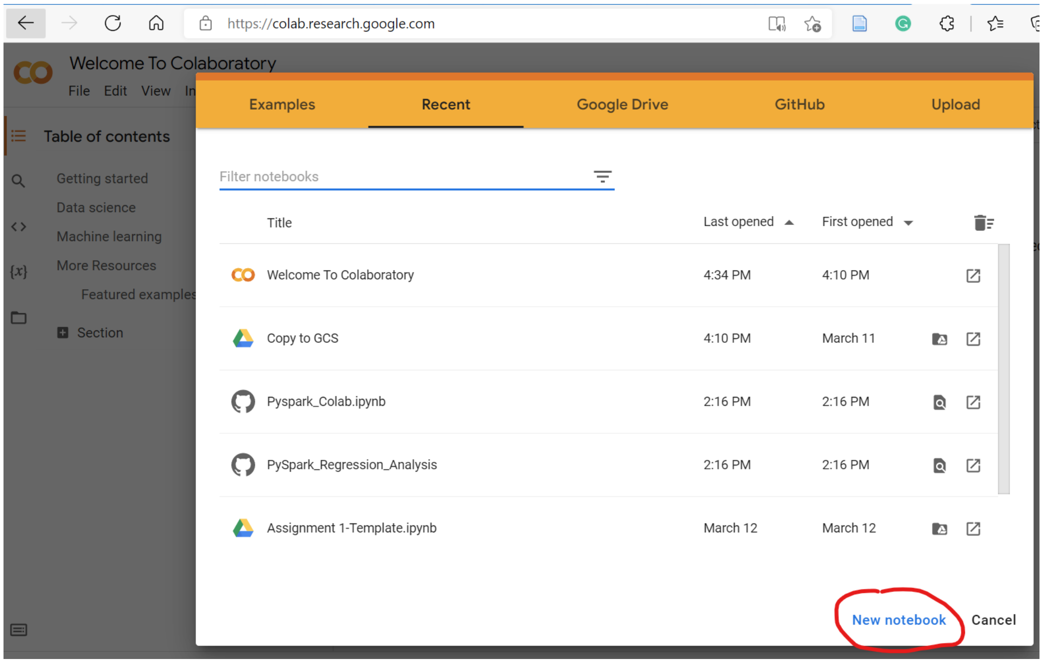
Using PySpark with Google Colab

Apache Spark (https://spark.apache.org/) is a multi-language engine for running data engineering, data science, and machine learning on a computer cluster or on a single machine.

PySpark (https://spark.apache.org/docs/latest/api/python/) is an interface for Apache Spark in Python. It allows writing Spark applications using Python APIs and provides the PySpark shell for interactively analyzing data in a distributed environment. PySpark supports most of Spark's features such as Spark SQL, DataFrame, Streaming, MLlib (Machine Learning), and Spark Core.

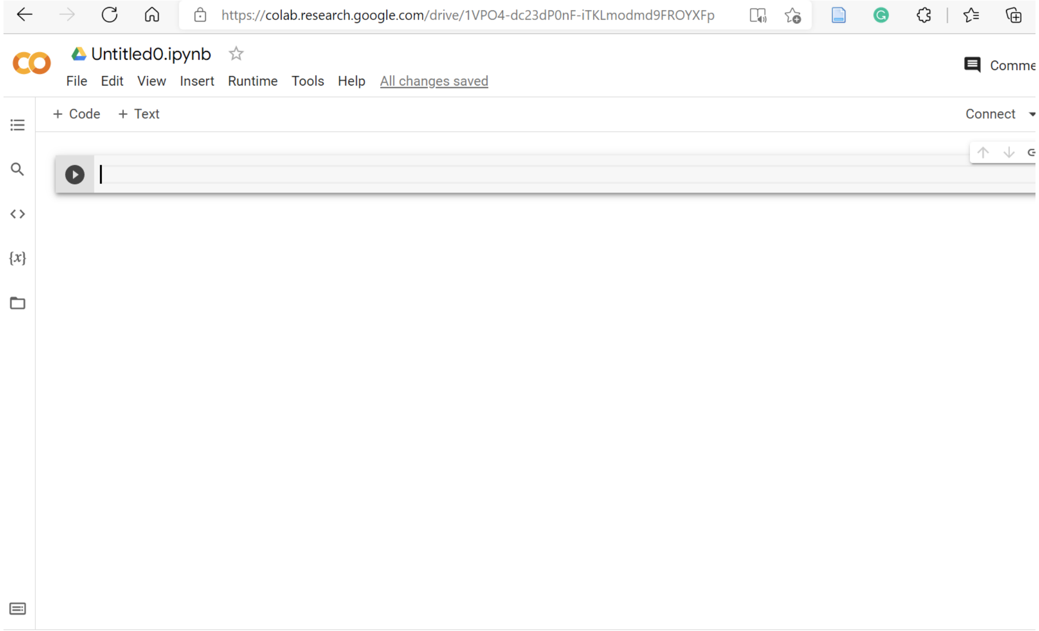
One of the easiest ways to use and try PySpark is through Google Colab. Colab is a product from Google Research and allows writing and executing Python code through the browser. Colab is a modified version of the Jupyter notebook hosted on Google servers and requires no setup while providing access free of charge to computing resources. It is particularly well suited to machine learning, data analysis, and education.

1. Go to https://colab.research.google.com/ and choose New notebook.



Title	Last opened	First opened
Welcome To Colaboratory	4:34 PM	4:10 PM
Copy to GCS	4:10 PM	March 11
Pyspark_Colab.ipynb	2:16 PM	2:16 PM
PySpark_Regression_Analysis	2:16 PM	2:16 PM
Assignment 1-Template.ipynb	March 12	March 12

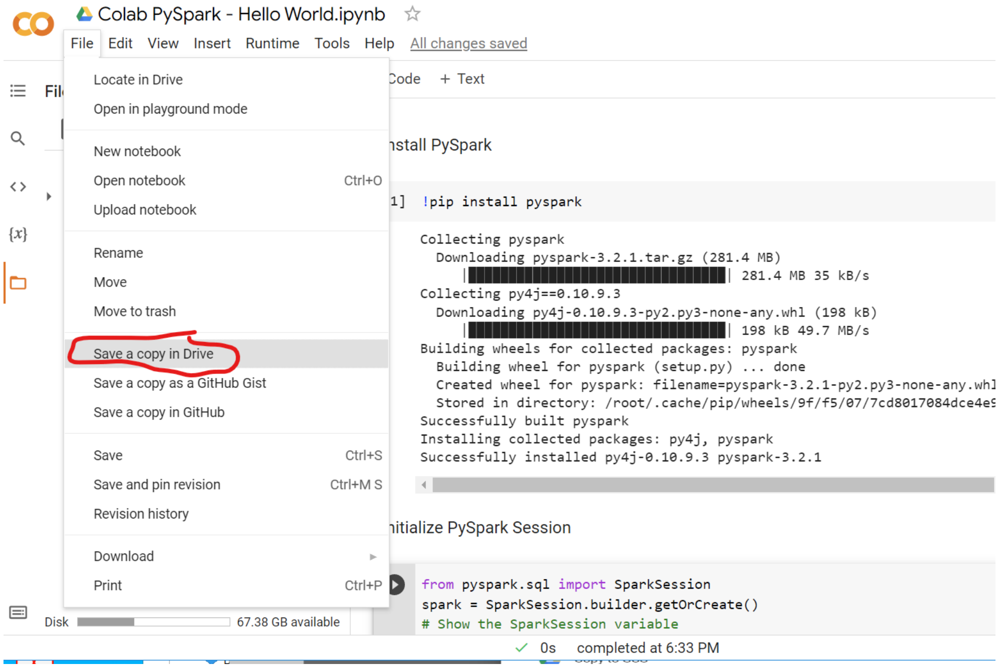
2. A new Jupiter notebook will be created, and you are ready to start writing your code.
3. You can click on the [+ Code] button to create new cells to enter your code.


https://colab.research.google.com/drive/1VPO4-dc23dP0nF-iTKLmodmd9FROYXFp
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comme
+ Code + Text
Connect

Example notebook

You can use the following example notebook to start using the PySpark.
https://colab.research.google.com/drive/1VPO4-dc23dP0nF-iTKLmodmd9FROYXFp?usp=sharing

1. When you open the notebook, you can create your own copy using Save a copy to Drive.

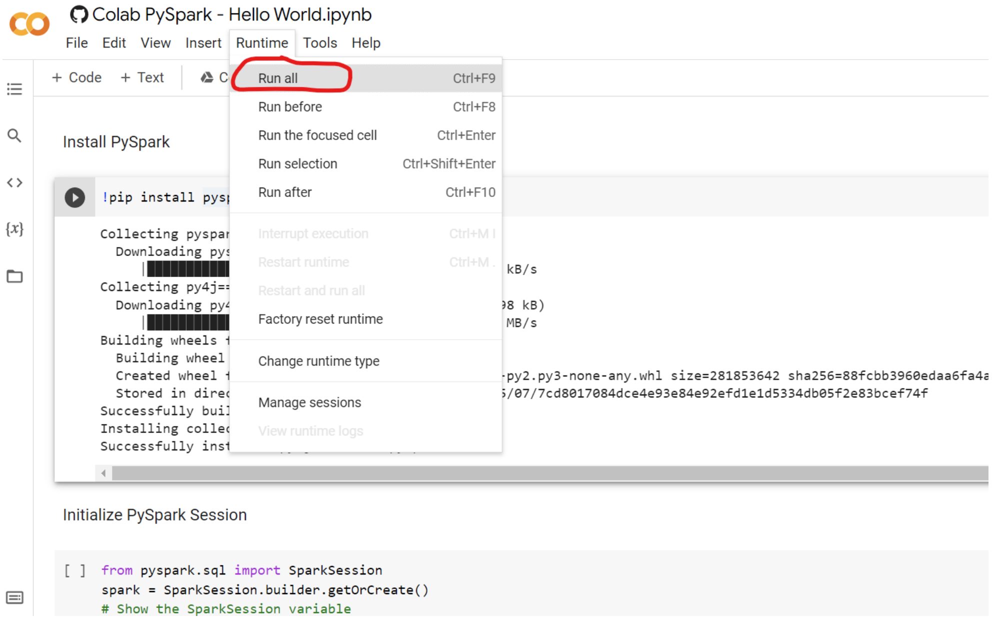


```
1]  !pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 35 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 49.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e9
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
# Show the SparkSession variable
```

2. Use the Run all command to run the notebook.



```
!pip install pys
```

```
Collecting pyspar                                   kB/s
  Downloading pys
     |████████
Collecting py4j==                                  98 kB)
  Downloading py4                                   MB/s
     |████████
Building wheels f
  Building wheel
  Created wheel f                          -py2.py3-none-any.whl size=281853642 sha256=88fcbb3960edaa6fa4a
  Stored in direc                          5/07/7cd8017084dce4e93e84e92efd1e1d5334db05f2e83bcef74f
Successfully buil
Installing collec
Successfully inst
```

```
[ ]  from pyspark.sql import SparkSession
     spark = SparkSession.builder.getOrCreate()
     # Show the SparkSession variable
```