



Search

Write



✦ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



Easy Install Pyspark in Anaconda

Install on Mac



Divya Chandana · Follow

Published in The AI Guide · 3 min read · Dec 11, 2022



89



2



Pre Installations:

Anaconda : <https://docs.anaconda.com/anaconda/install/mac-os/>

Step by Step to Install PySpark

1. *Install Java*
2. *Install pyspark*
3. *Install findspark*
4. *Test Spark Installation*
5. *Launch Anaconda*
6. *Launch Jupyter Lab*
7. *Spark Commands*

Step 1: Install Java

```
conda install openjdk
```

```
divyachandana — -bash — 108x45
(base) DCs-MacBook-Pro:~ divyachandana$ conda install openjdk
Collecting package metadata (current_repodata.json): done
Solving environment: done

==> WARNING: A newer version of conda exists. <==
  current version: 4.10.3
  latest version: 22.11.1

Please update conda by running

  $ conda update -n base -c defaults conda

## Package Plan ##

  environment location: /Users/divyachandana/opt/anaconda3

  added / updated specs:
    - openjdk

The following packages will be downloaded:



| package         | build      |          |
|-----------------|------------|----------|
| openjdk-11.0.13 | h8346a28_0 | 166.6 MB |
| Total:          |            | 166.6 MB |



The following NEW packages will be INSTALLED:

  openjdk                pkgs/main/osx-64::openjdk-11.0.13-h8346a28_0

Proceed ([y]/n)? y

Downloading and Extracting Packages
openjdk-11.0.13          | 166.6 MB | ##### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
(base) DCs-MacBook-Pro:~ divyachandana$
```

Step 2: Install pyspark

```
conda install pyspark
```

```
divvyachandana — python - -bash — 108x45
(base) DCs-MacBook-Pro:~ divvyachandana$ conda install pyspark
Collecting package metadata (current_repodata.json): done
Solving environment: failed with initial frozen solve. Retrying with flexible solve.
Solving environment: failed with repodata from current_repodata.json, will retry with next repodata source.
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /Users/divvyachandana/opt/anaconda3

  added / updated specs:
    - pyspark

The following packages will be downloaded:
```

package	build	
abseil-cpp-20210324.2	h23ab428_0	866 KB
arrow-cpp-3.0.0	py39ha6b1260_4	5.2 MB
aws-c-common-0.4.57	hb1e8313_1	140 KB
aws-c-event-stream-0.1.6	h23ab428_5	21 KB
aws-checksums-0.1.9	hb1e8313_0	47 KB
aws-sdk-cpp-1.8.185	he271ece_0	1.6 MB
boost-cpp-1.73.0	h9ed2024_11	25 KB
conda-4.12.0	py39hecd8cb5_0	14.5 MB
double-conversion-3.1.5	haf313ee_1	228 KB
gflags-2.2.2	h0a44026_0	104 KB
glog-0.5.0	h23ab428_0	87 KB
grpc-cpp-1.39.0	h8dc0301_5	2.2 MB
libboost-1.73.0	hd4c2dcd_11	12.7 MB
libevent-2.1.8	hddc9c9b_1	854 KB
libprotobuf-3.17.2	h2842e9f_1	1.8 MB
libthrift-0.14.2	h054ceb0_0	348 KB
orc-1.6.9	h2519b9b_3	339 KB
py4j-0.10.9.3	py39hecd8cb5_0	251 KB
pyarrow-3.0.0	py39hdf3e9eb_3	1.8 MB
pyspark-3.2.1	py39hecd8cb5_0	263.9 MB
re2-2022.04.01	he9d5cce_0	179 KB
uriparser-0.9.3	h0a44026_1	40 KB
utf8proc-2.6.1	h9ed2024_0	308 KB
Total:		307.5 MB

Step 3: Install findspark

```
conda install -c conda-forge findspark
```

```
divyachandana — -bash — 108x45
(base) DCs-MacBook-Pro:~ divyachandana$ conda install -c conda-forge findspark
Collecting package metadata (current_repodata.json): done
Solving environment: done

==> WARNING: A newer version of conda exists. <==
  current version: 4.12.0
  latest version: 22.11.1

Please update conda by running

  $ conda update -n base -c defaults conda

## Package Plan ##

  environment location: /Users/divyachandana/opt/anaconda3

  added / updated specs:
    - findspark

The following packages will be downloaded:



| package         | build          |         |             |
|-----------------|----------------|---------|-------------|
| conda-4.14.0    | py39h6e9494a_0 | 1022 KB | conda-forge |
| findspark-2.0.1 | pyhd8ed1ab_0   | 8 KB    | conda-forge |
| python_abi-3.9  | 2_cp39         | 4 KB    | conda-forge |
| Total:          |                | 1.0 MB  |             |



The following NEW packages will be INSTALLED:

  findspark      conda-forge/noarch::findspark-2.0.1-pyhd8ed1ab_0
  python_abi     conda-forge/osx-64::python_abi-3.9-2_cp39

The following packages will be UPDATED:

  conda          pkgs/main::conda-4.12.0-py39hecd8cb5_0 --> conda-forge::conda-4.14.0-py39h6e9494a_0


Proceed ([y]/n)? y
```

Step 4: Test Spark Installation

In command prompt

pyspark

```
[~]base) DCs-MacBook-Pro:~ divyachandana$ pyspark
Python 3.9.7 (default, Sep 16 2021, 08:50:36)
[Clang 10.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/Users/divyachandana/opt/anaconda3/lib/python3.9/site-packages/pyspark/jars/spark-unsafe_2.12-3.2.1.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/08 21:15:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/12/08 21:15:53 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to
```



version 3.2.1

```
Using Python version 3.9.7 (default, Sep 16 2021 08:50:36)
Spark context Web UI available at http://dcs-macbook-pro.home:4041
Spark context available as 'sc' (master = local[*], app id = local-1670555753796).
SparkSession available as 'spark'.
>>>
```

```

data_values = [('Apple',3),('Banana',6),('Orange', 9)]
column_name = ['Name', 'Count']
df = spark.createDataFrame(data_values).toDF(*column_name)
df.show()

```

Welcome to



version 3.2.1

Using Python version 3.9.7 (default, Sep 16 2021 08:50:36)

Spark context Web UI available at <http://dcs-macbook-pro.home:4041>

Spark context available as 'sc' (master = local[*], app id = local-1670555753796).

SparkSession available as 'spark'.

```
>>> data_values = [('Apple',3),('Banana',6),('Orange', 9)]
```

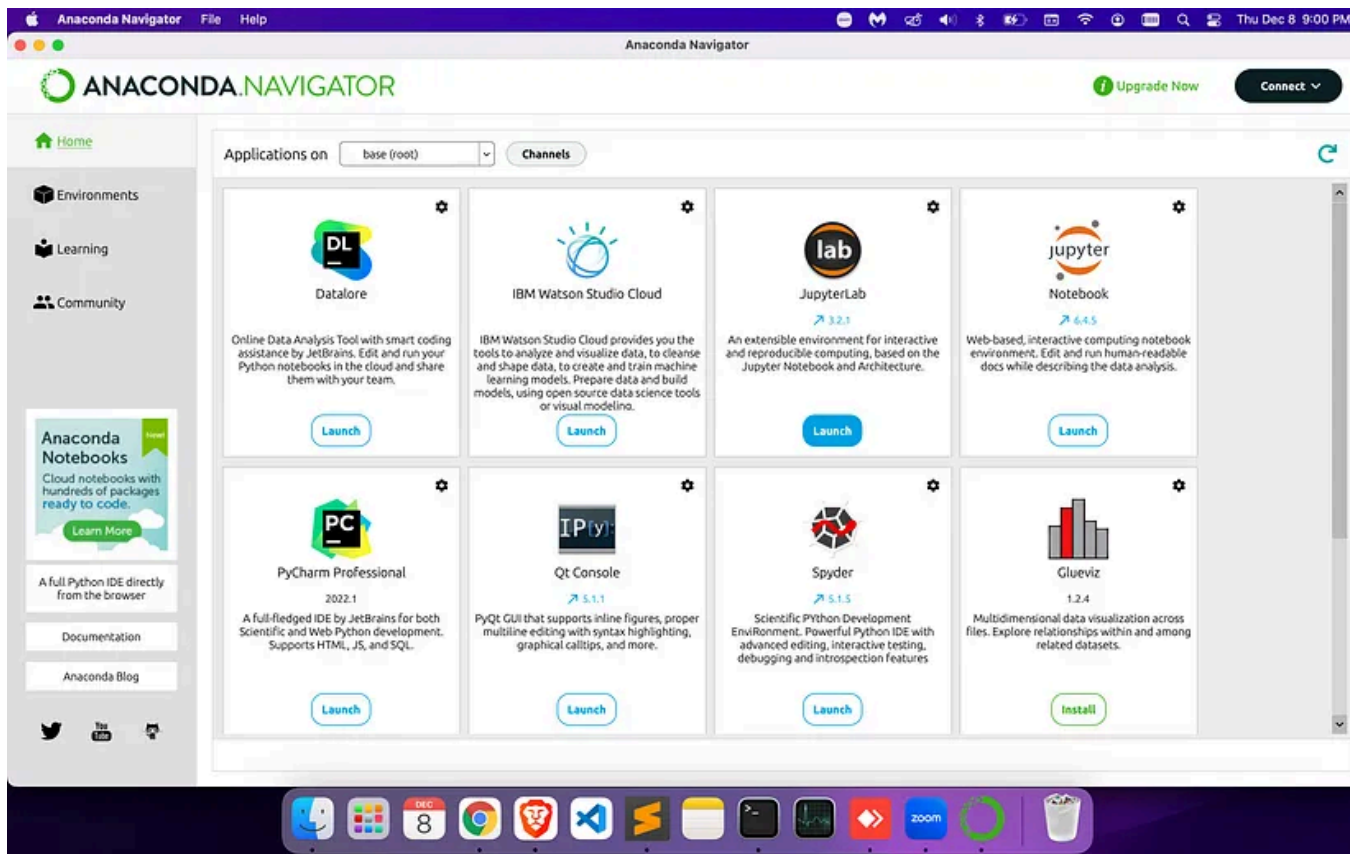
```
>>> column_name = ['Name', 'Count']
```

```
>>> df = spark.createDataFrame(data_values).toDF(*column_name)
```

```
>>> df.show()
```

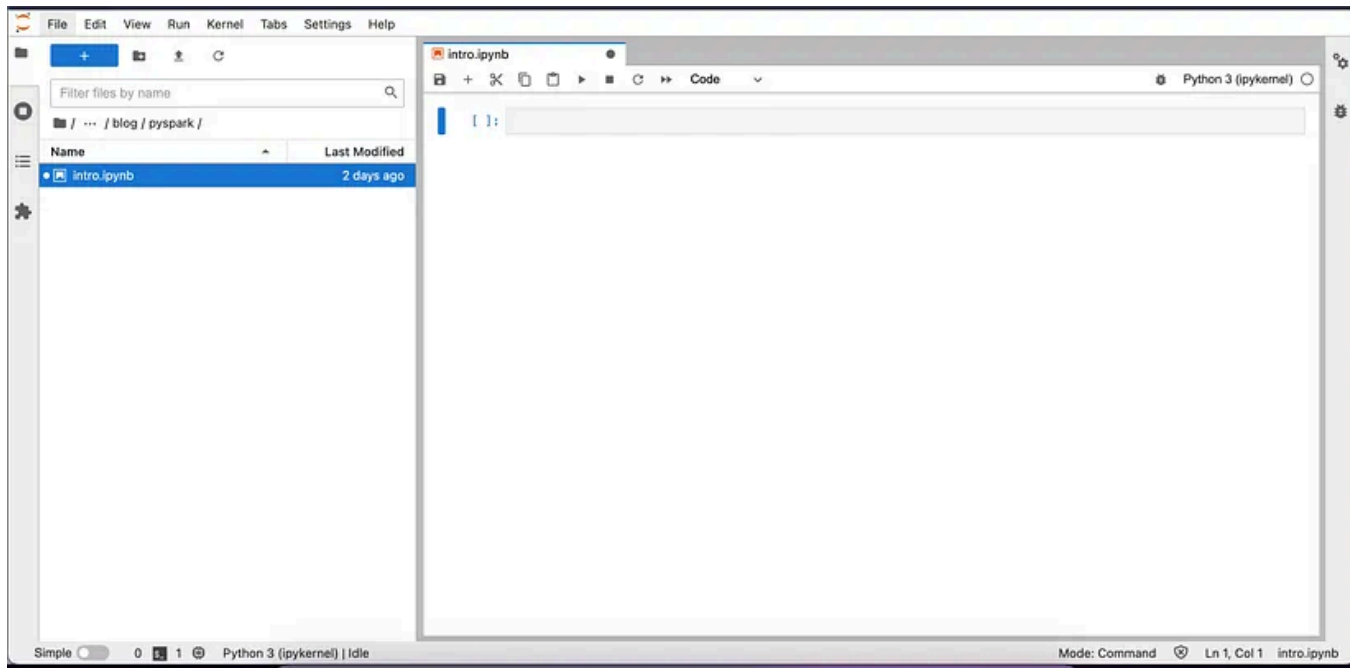
Name	Count
Apple	3
Banana	6
Orange	9

Step 5: Launch Anaconda



Step 6: Launch Jupyter Lab

<http://localhost:8888/lab>



Step 7: Spark Commands

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('pySparkSetup').getOrCreate()
```

```
[1]: from pyspark.sql import SparkSession

[2]: spark = SparkSession.builder.appName('pySparkSetup').getOrCreate()

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/Users/divyachandana/opt/anaconda3/lib/python3.9/site-packages/pyspark/jars/unsafe-3.2.1.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/08 21:06:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
data_values = [('Apple',3),('Banana',6),('Orange', 9)]
column_name = ['Name', 'Count']
df = spark.createDataFrame(data_values).toDF(*column_name)
df.show()
```

```
[3]: data_values = [('Apple',3),('Banana',6),('Orange', 9)]  
      column_name = ['Name', 'Count']
```

```
[5]: # create Spark Data Frame  
      df = spark.createDataFrame(data_values).toDF(*column_name)  
      df.show()
```

```
[Stage 0:> (0 + 1) / 1]
```

Name	Count
Apple	3
Banana	6
Orange	9

```
[ ]:
```

Stay tuned for next episodes for Pyspark

References

<https://sparkbyexamples.com/pyspark/install-pyspark-in-anaconda-jupyter-notebook/>

Introduction and Installation of PySpark for Mac Users

This blog briefly introduces you to the not-so-new but one of the most widely used technologies in Big Data and AI...

medium.com

How to Install PySpark and Apache Spark on MacOS - Luminis

Here is an easy Step by Step guide to installing PySpark and Apache Spark on MacOS. Step 1: Get Homebrew Homebrew...

www.luminis.eu

Complete Guide to Installing PySpark on MacOS

Getting PySpark set up locally can be a bit of an involved process that took me a few tries to get right. In this post...

kevinvecmanis.io

Pyspark

Jupyter Notebook

Python

Anaconda Navigator

Mac



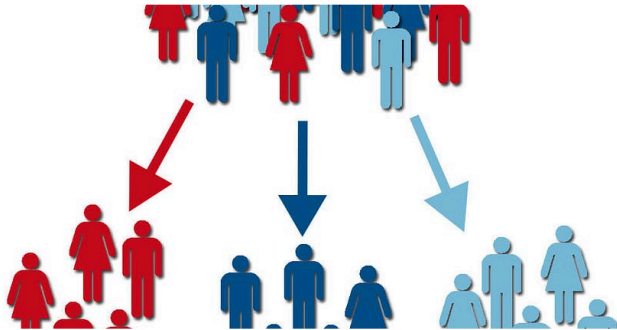
Written by Divya Chandana

75 Followers · Editor for The AI Guide

Follow



More from Divya Chandana and The AI Guide




 Divya Chandana

Exploring Customers Segmentation With RFM Analysis...

Introduction

9 min read · May 6, 2021



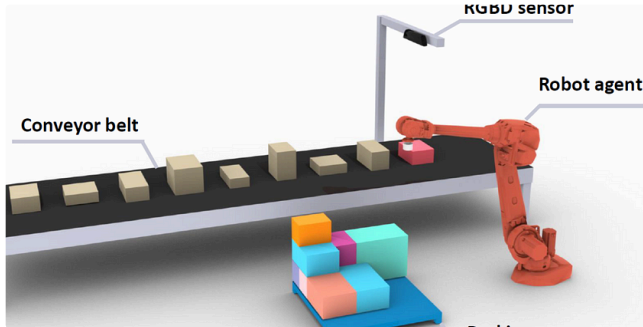
 Divya Chandana in The AI Guide


Stock Market Recommendation System

MOTIVATION

14 min read · May 12, 2021

👏 102 💬 1



 Divya Chandana in The AI Guide

Smart Automation bin Packing with Reinforcement Learning

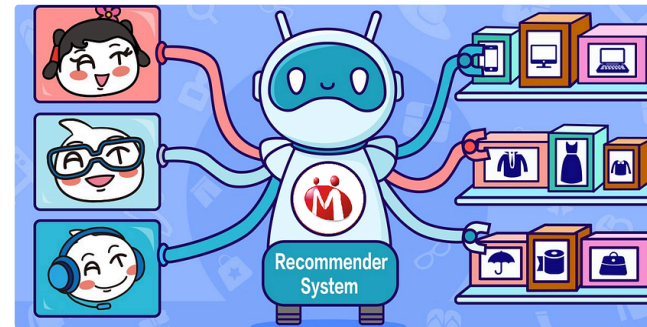
It's not about Kubernetes Automatic bin packing where it places containers...


6 min read · Mar 28, 2022

👏 40 💬 2



👏 114 💬 1



 Divya Chandana in The AI Guide

E-Commerce Recommendation Engine with Collaborative Filtering

Introduction

9 min read · May 7, 2021

👏 75 💬 2



See all from Divya Chandana

See all from The AI Guide

Recommended from Medium

Java 8 Java 8 Enterprise Performance Pack Java 11

Java SE Development Kit 8u391
Java SE subscribers will receive JDK 8 updates until at least **December 2030**.
Annual update required for some Java 8 users on macOS.
Oracle JDK 8 license changed in April 2019
Oracle Technology Network License Agreement for Oracle Java SE is substantially different from prior Oracle JDK 8 licenses. This license permits certain uses, such as personal use and cost -- but other uses authorized under prior Oracle JDK licenses may no longer be available. Please review the terms carefully before downloading and using this product. FAQs are available.
Commercial license and support are available for a low cost with Java SE Universal Subscription.
K 8 software is licensed under the Oracle Technology Network License Agreement for Oracle Java SE.
Java SE 8u391 checksums and DL 8 GPG Keys for RPMs

Linux macOS Solaris **Windows**

Product/file description	File size	Download
x86 Installer	139.66 MB	jdk-8u391-windows-i586.exe
x64 Installer	148.99 MB	jdk-8u391-windows-x64.exe

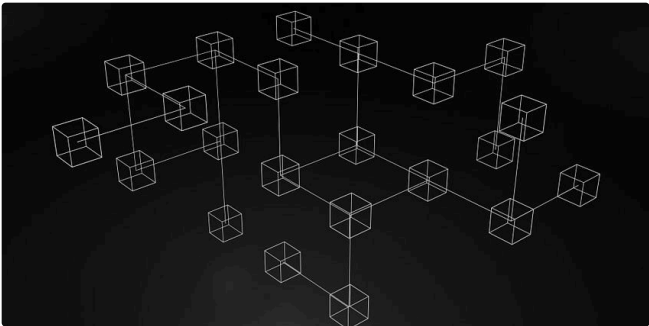
 Nidhi Gupta


Exploring PySpark Setup in Visual Studio Code

This article provides a step-by-step guide to setting up your environment, leveraging the...

3 min read · Dec 16, 2023

 135



 Adam Szpilewicz in Level Up Coding

Setting Up a Local Spark Playground with Docker Compose

In today's data-driven environment, Apache Spark stands out as an essential tool for big...

★ · 4 min read · Feb 1, 2024

 43



Lists



Coding & Development

11 stories · 510 saves



Predictive Modeling w/ Python

20 stories · 1011 saves



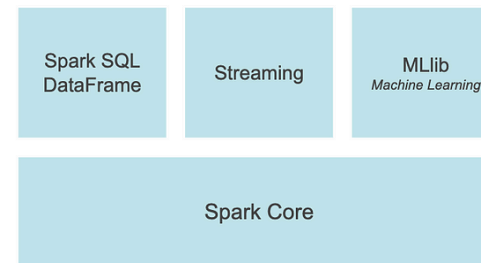
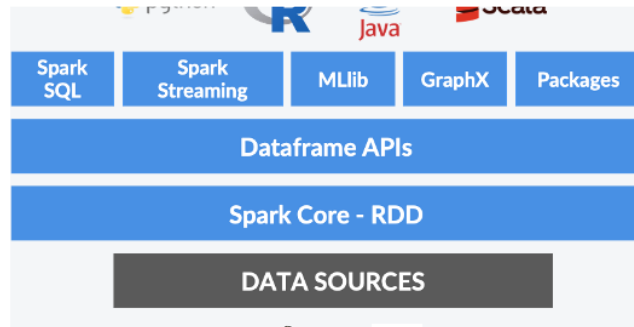
Practical Guides to Machine Learning

10 stories · 1211 saves



ChatGPT

21 stories · 524 saves



Sharan Harsoor in AI Mind

PySpark: Everything you need to know!

Welcome to this comprehensive guide on Apache Spark, a powerful distributed...

22 min read · Oct 27, 2023



Heber Bermudez

PySpark

Is an interface for Apache Spark in Python. It not only allows you to write Spark...

5 min read · Sep 28, 2023





 Nick Hass

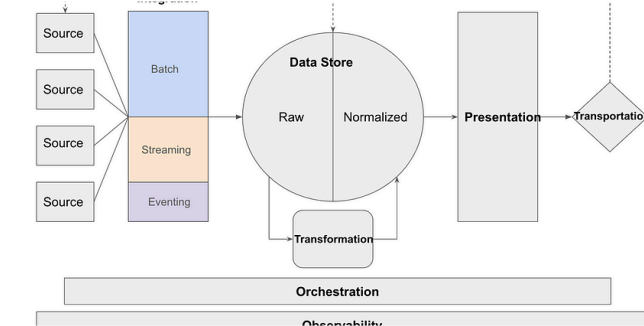
How to Connect Local PySpark to AWS S3 and Read a Delta Table

While you could use AWS EMR and automatically have access to the S3 file...

2 min read · Oct 4, 2023



5



 Dave Melillo in Towards Data Science

Building a Data Platform in 2024

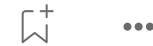
How to build a modern, scalable data platform to power your analytics and data...

9 min read · Feb 6, 2024



2.2K

34



See more recommendations