How do I install PySpark on Mac in 2024?

Authored by

stats writer

June 24, 2024

Installing PySpark on a Mac in 2024 is a straightforward process that requires a few simple steps. First, ensure that your Mac has the latest version of Python installed. Then, download and install the latest version of Apache Spark. Next, configure your environment variables to include the Spark and PySpark paths. Finally, install the PySpark library using the pip command. Following these steps will allow you to successfully install PySpark on your Mac, enabling you to use its powerful data processing capabilities for your Python projects.

Installing PySpark on macOS allows users to experience the power of Apache Spark, a distributed computing framework, for big data processing and analysis using Python. PySpark seamlessly integrates Spark's capabilities with Python's simplicity and flexibility, making it an ideal choice for data engineers and data scientists working on large-scale data projects.

To install PySpark on macOS, users typically follow a series of steps that involve setting up the Java Development Kit (JDK), installing Apache Spark, configuring Python, and setting environment variables. Additionally, installing the findspark package can streamline the process by facilitating the location of the Spark installation within Python scripts.

PySpark installation steps for Mac OS using Homebrew

Related Articles

1. Install PySpark on Mac using Homebrew

Homebrew is a package manager for macOS and Linux systems. It allows users to easily install, update, and manage software packages from the command line. With Homebrew, users can install a wide range of software packages and utilities, including development tools, programming languages, libraries, and applications, directly from the terminal.

To use homebrew, first you need to install it.

```
# Install Homebrew
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

This command will prompt for the root password. You'll be required to enter your root password to execute this command. On a personal laptop, this password is the same as the one used when logging into your Mac. If you lack root access, reach out to your system administrator. Upon successful installation of Homebrew, you should see a message similar to the following.

```
==> Installation successful!

==> Homebrew has enabled anonymous aggregate formulae and cask analytics.
Read the analytics documentation (and how to opt-out) here:
  https://docs.brew.sh/Analytics
No analytics data has been sent yet (nor will any be during this install run).

==> Homebrew is run entirely by unpaid volunteers. Please consider donating:
  https://github.com/Homebrew/brew#donations

==> Next steps:
- Run these two commands in your terminal to add Homebrew to your PATH:
    echo 'eval "$(/opt/homebrew/bin/brew shellenv)"' >> /Users/admin/.zprofile
    eval "$(/opt/homebrew/bin/brew shellenv)"
- Run brew help to get started
- Further documentation:
    https://docs.brew.sh
```

Once the installation is done, set the homebrew to your `$PATH` environment variable by using the below command.

```
# Set brew to Path
echo 'eval "$(/opt/homebrew/bin/brew shellenv)"' >> /Users/admin/.zprofile
eval "$(/opt/homebrew/bin/brew shellenv)"
```

If you have issues with the above process, follow the instructions from Homebrew to install it.

Note: When users interact with Homebrew from the terminal, they typically use commands like `brew install`, `brew update`, or `brew upgrade` to manage software installations and updates. These commands are part of the Homebrew package manager

2. Install Java Development Kit (JDK)

Java is a prerequisite for running PySpark as it provides the runtime environment necessary for executing Spark applications. When PySpark is initialized, it starts a JVM (Java Virtual Machine) process to run the Spark runtime, which includes the Spark Core, SQL, Streaming, MLlib, and GraphX libraries. This JVM process executes the Spark code.

Java from Oracle is not open-source hence, I will use Java from openjdk and use brew to install it. The following command install Java/JDK 11 version from openjdk.

```
# Install OpenJDK 11
brew install openjdk@11
```

Note: You need to install a Java version that is compatible with the Apache Spark/PySpark you going to install.

3. Install Python

PySpark is a Python library; hence, you need Python to run.

3.1 With Virtual Environment (Recommended)

MacOS, by default, comes with a Python version, and it is recommended not to touch that version as it is needed to run several Mac applications. Hence, I will create a virtual environment and install the required Python version.

```
brew install pyenv # Install pyenv
pyenv install 3.11.5 # Install Python version
brew install pyenv-virtualenv # Required to create a virtual environment
pyenv virtualenv 3.11.5 devenv # Create virtual environment devenv with python version 3.11.5
pyenv shell devenv # Initialize virtualenv for your shell
```

To activate and use the `devenv` virtual environment, you need to run the following command every time when you open a new terminal.

```
# Activate devenv virtual environment
pyenv shell devenv
```

3.2 Without Virtual Environment

Using the brew command, install Python without a virtual environment.

```
# Install Python
brew install python
```

Note: You need to install a Python version that is compatible with the Apache Spark/PySpark you going to install.

4. Install PySpark Latest Version on Mac

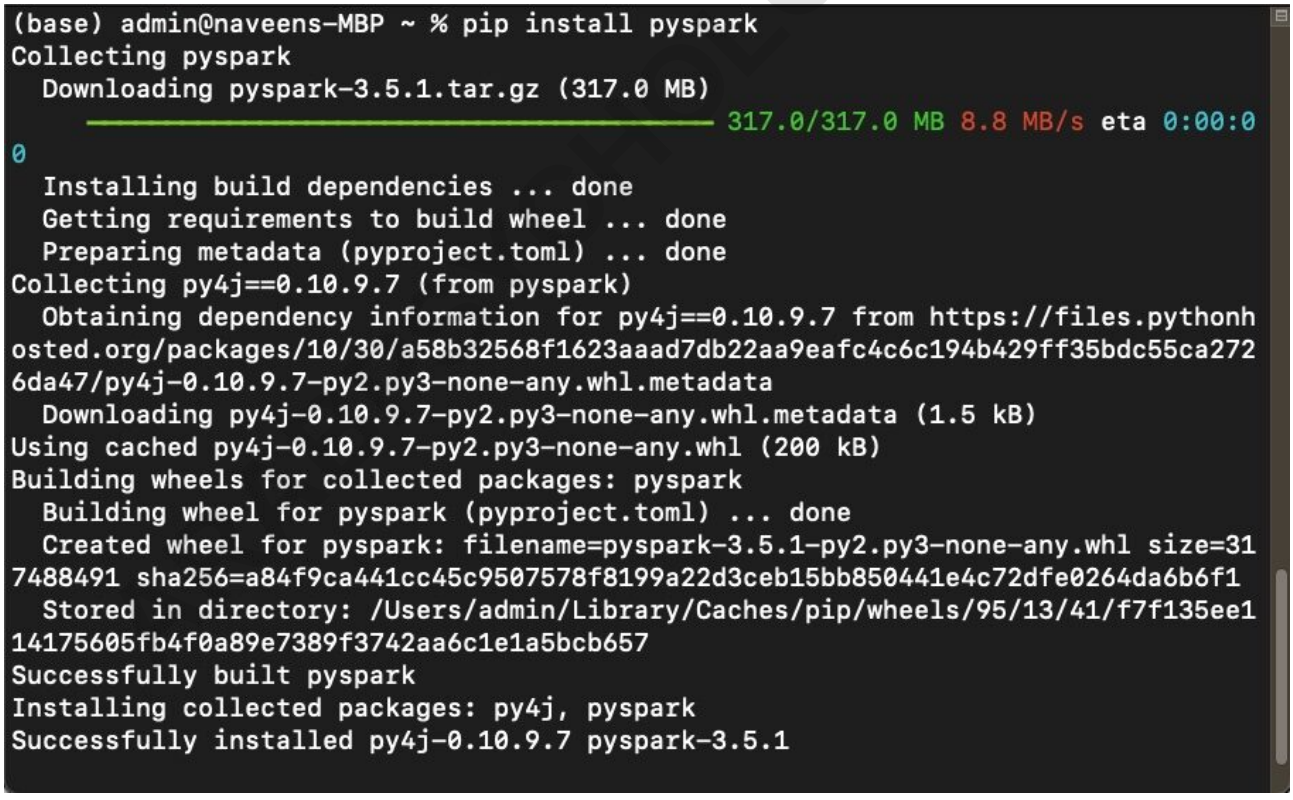
PySpark is available in PyPI, so it is easy to install from here. Installing PySpark via pip (the PyPI package manager) is straightforward and can be done with a single command, eliminating the need for manual downloads and configurations.

PyPI manages dependencies automatically, ensuring that all required packages and dependencies are installed correctly, saving time and effort.

To install PySpark from PyPI, you should use the pip command.

```
# Install Python
pip install pyspark
```

You should see something like the below



install pyspark using pip

Alternatively, you can also install Apache Spark using the brew command.

```
# Install Apache Spark
```

```
brew install apache-spark
```

5. Set Environment Variables

If you installed Apache Spark instead of PySpark, you need to set the `SPARK_HOME` environment variable to point to the directory where Apache Spark is installed.

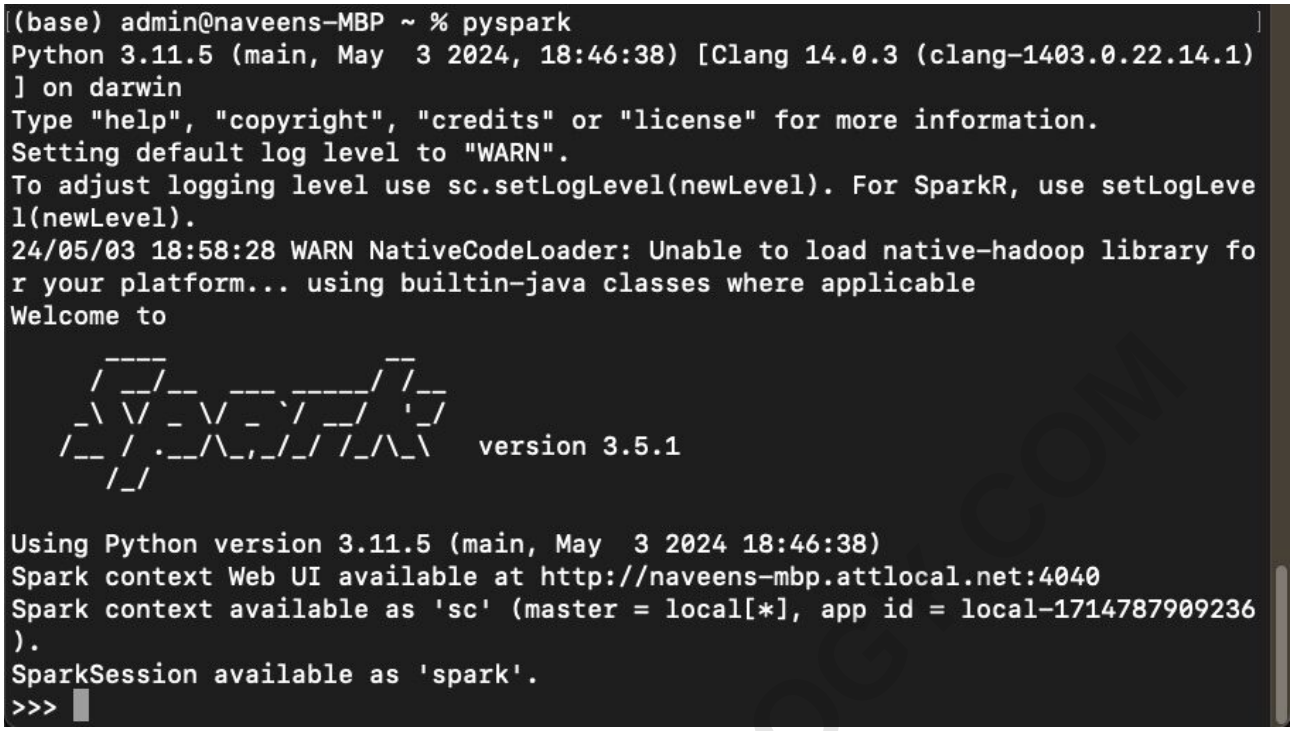
And, you also need to set the `PYSPARK_PYTHON` environment variable to point to your Python executable, typically located at `/usr/local/bin/python3`.

Setting the `PYSPARK_PYTHON` environment variable is important when working with PySpark because it allows users to specify which Python executable should be used by PySpark. This is particularly useful in environments where multiple versions of Python are installed or when PySpark needs to run with a specific Python interpreter.

6. Validate PySpark Installation from Shell

Once the PySpark or Apache Spark installation is done, start the PySpark shell from the command line by issuing the `pyspark` coammand.

The PySpark shell refers to the interactive Python shell provided by PySpark, which allows users to interactively run PySpark code and execute Spark operations in real-time. It provides an interactive environment for exploring and analyzing data using PySpark without the need to write full Python scripts or Spark applications.



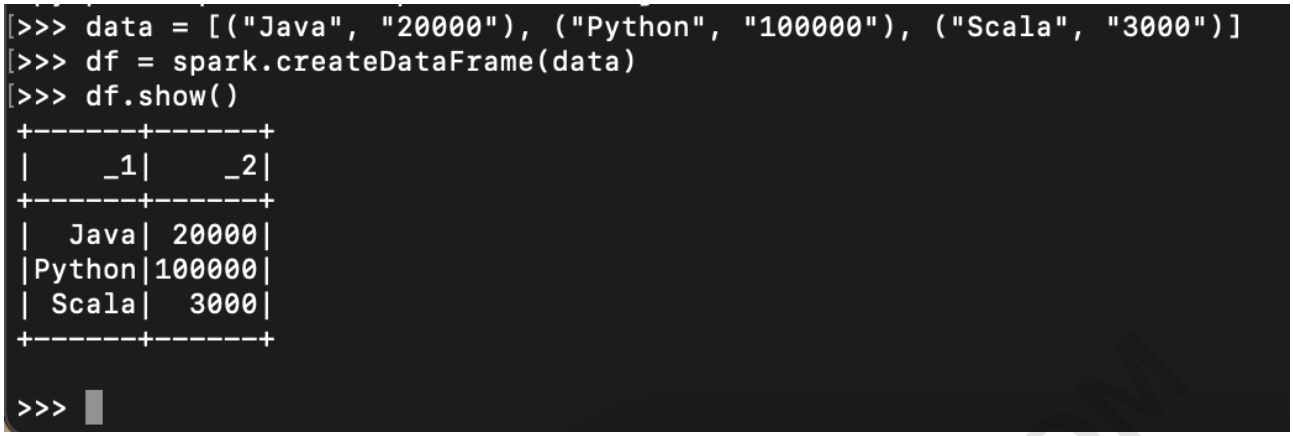
pyspark shell

7. Initiate DataFrame

Finally, let's create a DataFrame to confirm the installation is done successfully.

```
# Create DataFrame in PySpark Shell
data =
df = spark.createDataFrame(data)
df.show()
```

Yields below output.



For more examples on PySpark, refer to PySpark Tutorial with Examples.

Conclusion

In conclusion, installing PySpark on macOS is a straightforward process that empowers users to leverage the powerful capabilities of Apache Spark for big data processing and analytics. I hope you have set up PySpark on your macOS systems by following the installation steps.

Happy Learning !!

Related Articles