How can I install PySpark on Windows?

Authored by

stats writer

June 24, 2024

PySpark is a powerful open-source tool used for data analysis and processing in Python. It is widely used in various industries for its scalability and performance. To install PySpark on a Windows operating system, there are a few steps that need to be followed. Firstly, the latest version of Python needs to be installed. Then, the Apache Spark software needs to be downloaded and extracted. After that, the necessary environment variables need to be set. Finally, the PySpark package needs to be installed using the pip command. These steps will allow users to successfully install PySpark on their Windows system and take advantage of its features for their data analysis needs.

In this article, we'll focus specifically on how to install PySpark on the Windows operating system. While Spark is primarily designed for Unix-based systems, setting it up on Windows can sometimes be a bit tricky due to differences in environment and dependencies. However, with the right steps and understanding, you can install PySpark into your Windows environment and run some examples.

I will also cover how to start a history server and monitor your jobs using Web UI.

To **Install PySpark on Windows** follow the below step-by-step instructions.

Install Python or Anaconda distribution

Download and install either Python from Python.org or Anaconda distribution which includes Python, Spyder IDE, and Jupyter Notebook. I would recommend using Anaconda as it's popular and used by the Machine Learning and Data science community.

To use Anaconda distribution, follow Install PySpark using Anaconda & run Jupyter notebook

Install Java 8

To run the PySpark application, you would need Java 8/11/17 or a later version. Download and install JDK from OpenJDK.

Once the installation completes, set JAVA_HOME and PATH variables as shown below. Change the JDK path according to your installation.

```
JAVA_HOME = C:Program FilesJavajdk1.8.0_201
PATH = %PATH%;C:Program FilesJavajdk1.8.0_201bin
```

PySpark Install on Windows

You can install PySpark either by downloading binaries from spark.apache.org or by using the Python pip command.

Install using Python PiP

Python pip, short for "Python Package Installer," is a command-line tool used to install, manage, and uninstall Python packages from the Python Package Index (PyPI) or other package indexes. PyPI is a repository of software packages developed and shared by the Python community.

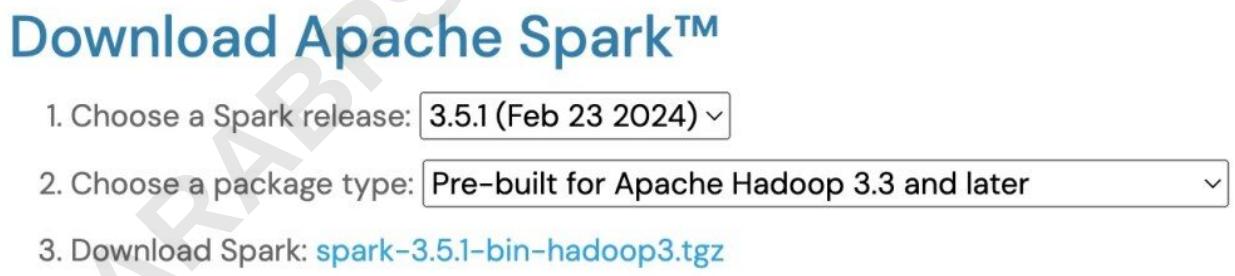
PySpark is available in PyPI hence, you can install it using the pip command.

```
# Install pyspark using pip command
pip install pyspark
```

Download & Install from spark.apache.org

If you install PySpark using PIP, then skip this section.

Access the Spark Download page, choose the Spark release version and package type; the link on point 3 updates to the selected options. select the link to download it.



Screenshot

2. Unzip the binary using WinZip or 7zip and copy the underlying folder `spark-3.5.1-bin-hadoop3` to `c:apps`

3. Open the Windows environment setup screen and set the following environment variables.

```
SPARK_HOME = C:appsspark-3.5.1-bin-hadoop3
HADOOP_HOME = C:appsspark-3.5.1-bin-hadoop3
```

```
PATH=%PATH%;C:appsspark-3.5.1-bin-hadoop3bin
```

Install winutils.exe on Windows

winutils.exe is a set of utilities for Windows used in Hadoop deployments. These utilities are primarily required for running Apache Hadoop applications on a Windows operating system. Copy winutils files to `%SPARK_HOME%bin` folder.

PySpark shell

The PySpark shell is an interactive Python shell that provides a convenient way to interact with Apache Spark. To launch the PySpark shell, you typically use the `pyspark` command in your terminal or command prompt. Once launched, you'll see the Python interpreter prompt (`>>>`) indicating that you can start executing Python code. From there, you can import the `pyspark` module and start interacting with Spark.

```
(base) admin@naveens-MBP ~ % pyspark
Python 3.11.5 (main, May  3 2024, 18:46:38) [Clang 14.0.3 (clang-1403.0.22.14.1)
] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/05/03 18:58:28 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/

Using Python version 3.11.5 (main, May  3 2024 18:46:38)
Spark context Web UI available at http://naveens-mbp.attlocal.net:4040
Spark context available as 'sc' (master = local[*], app id = local-1714787909236
).
SparkSession available as 'spark'.
>>>
```

Screenshot

Run the below statements in PySpark shell to create an RDD.

```
# RDD creation
rdd = spark.sparkContext.parallelize()
```

```
print(rdd.count)
```

Spark-shell generates a Spark context web UI, which is accessible by default at http://localhost:4040.

Web UI

The Spark Web UI or Spark UI, is a web-based interface provided by Apache Spark for monitoring and managing Spark applications. It offers real-time insights into the execution of Spark jobs, providing information about tasks, stages, executors, and more.

You can access Spark Web UI by accessing http://localhost:4040. You can find this URL on the PySpark shell console.

Conclusion

In summary, you have learned how to install PySpark on Windows and run sample statements in spark-shell. If you have any issues setting it up, please message me in the comments section, and I will try to respond with a solution.

Happy Learning !!

Related Articles