Install Apache Spark on Windows 10

# Overview

# We cover installation, testing, and initial exploration of Apache Spark on Windows 10. This document describes installing Spark 3.4.1 to be used with Java 17 and Python 3.11.4. We also describe tests and initial Spark observability tools.

# Prerequisites

Spark installation requires you to:

* There are: 4 GM of RAM and 10 GB of free disk space available.
* Download Spark and winutils tools.
* Create directories and place software in them.
* Create and define Windows environment variables.
* If needed, install compatible versions of Java and Python (URLs below help define compatible versions.)

We presume that Java 17 is installed and available, and Python 3.8 or better is installed. The environment variable JAVA\_HOME must be defined and point to the correct Java runtime. The Python executable needs to be on the PATH.

# Sample Downloaded Spark and Winutils Resources

Winutils details are discussed below. Once assembled, you should have the following files (or similar choices):

D:\Temp2>**dir/s spark\_support\_tools**

Directory of D:\Temp2\spark\_support\_tools

08/04/2023 06:11 PM 397,281,767 spark-3.4.1-bin-hadoop3-scala2.13.tgz

08/04/2023 06:07 PM <DIR> winutils\_3\_3\_5

1 File(s) 397,281,767 bytes

Directory of D:\Temp2\spark\_support\_tools\winutils\_3\_3\_5

08/04/2023 05:55 PM 84,992 hadoop.dll

08/04/2023 05:57 PM 38,622 hadoop.lib

08/04/2023 05:57 PM 1,246,294 libwinutils.lib

08/04/2023 06:07 PM 112,640 winutils.exe

# Verify Java and Python Versions

D:\Temp2>**set JAVA**

JAVA17\_HOME="D:\Java\jdk-17"

JAVA20\_HOME="D:\Java\jdk-20"

JAVA\_HOME="D:\Java\jdk-20"

JAVA\_OPTIONS=-Xmx512M -Xms512M

D:\Temp2>**set JAVA\_HOME=%JAVA17\_HOME%**

D:\Temp2>**%JAVA\_HOME%\bin\java -version**

java version "17.0.6" 2023-01-17 LTS

Java(TM) SE Runtime Environment (build 17.0.6+9-LTS-190)

Java HotSpot(TM) 64-Bit Server VM (build 17.0.6+9-LTS-190, mixed mode, sharing)

D:\Temp2>**python**

Python 3.11.4 (tags/v3.11.4:d2340ef, Jun 7 2023, 05:45:37) [MSC v.1934 64 bit (AMD64)] on win32

Type "help", "copyright", "credits" or "license" for more information.

>>> quit()

# Installation Steps

# Install basic Spark software

We first install the Windows version of Spark as follows:

1. Download the latest version of the Spark executable compressed file from the link given below to a temporary working directory.
2. Extract the required software with 7zip; two intermediates, tgz to tar to the Spark directory.
3. Move the final extracted software to the intended Spark “home” directory.
4. Spark will look like this:

D:\util>**dir spark-3.4.1-bin-hadoop3-scala2.13**

08/04/2023 06:17 PM <DIR> .

08/04/2023 06:17 PM <DIR> ..

06/19/2023 03:39 PM <DIR> bin

06/19/2023 03:39 PM <DIR> conf

06/19/2023 03:39 PM <DIR> data

06/19/2023 03:39 PM <DIR> examples

06/19/2023 03:39 PM <DIR> jars

06/19/2023 03:39 PM <DIR> kubernetes

06/19/2023 03:39 PM 22,982 LICENSE

06/19/2023 03:39 PM <DIR> licenses

06/19/2023 03:39 PM 57,842 NOTICE

06/19/2023 03:39 PM <DIR> python

06/19/2023 03:39 PM <DIR> R

06/19/2023 03:39 PM 4,605 README.md

06/19/2023 03:39 PM 165 RELEASE

06/19/2023 03:39 PM <DIR> sbin

06/19/2023 03:39 PM <DIR> yarn

# Install Winutils Tools

We need to create a “HADOOP home” holding the winutils utilities files. We were able to get Spark working with just winutils.exe in our HADOOP home directory, and did not require the DLL and LIB files. The resulting installation should look like this:

D:\util>**dir/s Hadoop**

Directory of D:\util\Hadoop\winutils\_3\_3\_5\bin

08/04/2023 06:07 PM 112,640 winutils.exe

# Optional Spark “Scratch Directory” Definition

Spark uses the local file system for temporary file creation. This work area typically defaults to the C drive. Many of use another drive for working storage. We show how to change the Spark working storage.

Copy file spark-defaults.conf.template and rename it spark-defaults.conf in the Spark conf directory. Add a line to the end of the file similar to this example:

# spark.driver.memory 5g

# spark.executor.extraJavaOptions -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"

***spark.local.dir******D:\Temp***

In addition, define this environment variable for your situation: *SPARK\_LOCAL\_DIRS=D:\Temp*.

# Required Environment Variables

Spark and Hadoop tools are located for runtime use with these environment variables that you need to define:

HADOOP\_HOME=D:\util\Hadoop\winutils\_3\_3\_5

SPARK\_HOME=D:\util\spark-3.4.1-bin-hadoop3-scala2.13

PYSPARK\_DRIVER\_PYTHON=python

PYSPARK\_PYTHON=python

# Start PySpark

Executing the Python shell, you will see something like the test below. Terminate the shell with control-z and return.

D:\Temp2**>%SPARK\_HOME%\bin\pyspark**

Python 3.11.4 (tags/v3.11.4:d2340ef, Jun 7 2023, 05:45:37) [MSC v.1934 64 bit (AMD64)] on win32

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 setLogLevel(newLevel).

23/08/05 17:38:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

23/08/05 17:38:40 WARN SparkConf: Note that spark.local.dir will be overridden by the value set by the cluster manager (via SPARK\_LOCAL\_DIRS in mesos/standalone/kubernetes and LOCAL\_DIRS in YARN).

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_ / .\_\_/\\_,\_/\_/ /\_/\\_\ version 3.4.1

/\_/

Using Python version 3.11.4 (tags/v3.11.4:d2340ef, Jun 7 2023 05:45:37)

Spark context Web UI available at **http://THOR.mshome.net:4040**

Spark context available as 'sc' (master = local[\*], app id = local-1691282323615).

SparkSession available as 'spark'.

>>> print(spark.version)

3.4.1

>>> ^Z[RETURN]

D:\Temp2>SUCCESS: The process with PID 2524 (child process of PID 21044) has been terminated.

SUCCESS: The process with PID 21044 (child process of PID 20736) has been terminated.

SUCCESS: The process with PID 20736 (child process of PID 12384) has been terminated.

***Important Note***: The **red** URL above is an observation entry into the Spark runtime called “the Web UI”. Record this URL for future use. Also note that the generic URL <http://localhost:4040> also works.

#### Optional Scratch Work Area Verification

For our example, we verify that the Spark work area is in D:\Temp:

Directory of D:\Temp

08/05/2023 05:38 PM <DIR> blockmgr-3e851b88-ddb8-4a52-bd5a-e6fd6bfaae47

08/05/2023 05:38 PM <DIR> spark-0b3ad36b-0f53-4431-b864-06874ae8dc47

# Start Spark-Shell

We can use the spark-shell as well. Skipping the Spark header, you should see something like this:

scala> spark.version

val res0: String = 3.4.1

scala> val df = sc.parallelize(Array(1,2,3,4,5,6,7,8,9,10)).toDF()

warning: 1 deprecation (since 2.13.0); for details, enable `:setting -deprecation` or `:replay -deprecation`

val df: org.apache.spark.sql.DataFrame = [value: int]

scala> df.printSchema()

root

|-- value: integer (nullable = false)

scala> df.show()

+-----+

|value|

+-----+

| 1|

| 2|

| 3|

| 4|

| 5|

| 6|

| 7|

| 8|

| 9|

| 10|

+-----+

Terminate the spark-shell with control-d.

# Installation Validation

# Ddd

# Ddd

# Ddd

# Ddd

# dd

# Winutils Background

The “winutils.exe” utility provides POSIX command support for Hadoop code and is used my Spark as well. This utility, along with the supporting libraries and DLL files, must be provided for the Windows 10 operating system. The adventurous may wish to build their own winutils artifacts on Windows with C++; and they may reference the *2021-04-14 – Build your own WINUTILS* article below. This article defines the winutils dependencies you need to create if you have security issues: build of your own winutils.exe, libwinutils.lib, hadoop.dll and hadoop.lib from known source code.

# Dependencies

# Runtime Dependencies

* **Spark Versioning Policy**: <https://spark.apache.org/versioning-policy.html>
* **Spark 3.4 Dependencies**: <https://spark.apache.org/docs/latest/>
* **Historic Spark Version Differences:** <https://www.educba.com/spark-versions/>

|  |  |  |
| --- | --- | --- |
| **Language** | **Supported** | **Deprecated** |
| Spark Runtime - JVM | 8/11/17 | 8 |
| R | 3.5+ |  |
| Python | 3.7+ | 3.7 |
| Scala | 2.12/2.13 | Scala App compiled version matches Spark compiled version |

# Spark Downloads and Utilities

* Spark Distribution Downloads: <https://spark.apache.org/downloads.html>
* *Original* WINDUTILS site: <https://github.com/steveloughran/winutils>
* *New* WINUTILS site: <https://github.com/cdarlint/winutils>
* 2021-04-14 – Build your own WINUTILS: <https://pivotalbi.com/build-your-own-winutils-for-spark/>

# References

Spark Usage Resources:

* Configuration Guide: <https://spark.apache.org/docs/latest/configuration.html>.

Additional Spark Installation Instructions:

* 2023-06-14: <https://www.knowledgehut.com/blog/big-data/how-to-install-apache-spark-on-windows> (v3.4.0)
* 2023-01-27: <https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/> (v3.0.0)
* 2021-Spark 3.1.2: <https://denisecase.github.io/starting-spark/>

# Sss

# Ssss

# Sss

# s