Install Apache Spark on Windows 10

# Overview

# We cover installation, testing, and initial exploration of Apache Spark on Windows 10.

# Prerequisites

# Sample Downloaded Resources

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

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

08/04/2023 11:07 PM <DIR> .

08/04/2023 11:07 PM <DIR> ..

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 06:07 PM <DIR> .

08/04/2023 06:07 PM <DIR> ..

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

# Dd

Setup the winutils

# Dd

Setup the environment variables

# d

# 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 from the link given below to a temporary working directory.
2. Extract the required software with 7zip; two intermediates, tgz to tar to directory.
3. Move the final extracted software to the intended Spark “home” directory.
4. It 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

# Sss

# Sss

# Sss

# sps

# xxxtwo

# dddf

# Hive Workspace

The Hive metadata work space location is controlled by configuration described [here](https://cwiki.apache.org/confluence/display/Hive/Configuration+Properties%23ConfigurationProperties-hive.exec.scratchdir).

# Winutils

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