Kontext



Install Apache Spark 3.0.0 on Windows 10



Raymond ⊚ 9,782 ■ 8 🖶 2020-08-09 🕓 3 years ago

Kontext Big Data Tools Apache Spark

Spark 3.0.0 was release on 18th June 2020 with many new features. The highlights of features include adaptive query execution, dynamic partition pruning, ANSI SQL compliance, significant improvements in pandas APIs, new UI for structured streaming, up to 40x speedups for calling R user-defined functions, accelerator-aware scheduler and SQL reference documentation.

This article summarizes the steps to install Spark 3.0 on your Windows 10 environment.

Tools and Environment

- GIT Bash
- Command Prompt
- Windows 10
- Python
- Java JDK

Install Git Bash

Download the latest Git Bash tool from this page: https://git-scm.com/downloads.

Run the installation wizard to complete the installation.

Install Java JDK

Spark 3.0 runs on Java 8/11. You can install Java JDK 8 based on the following section.

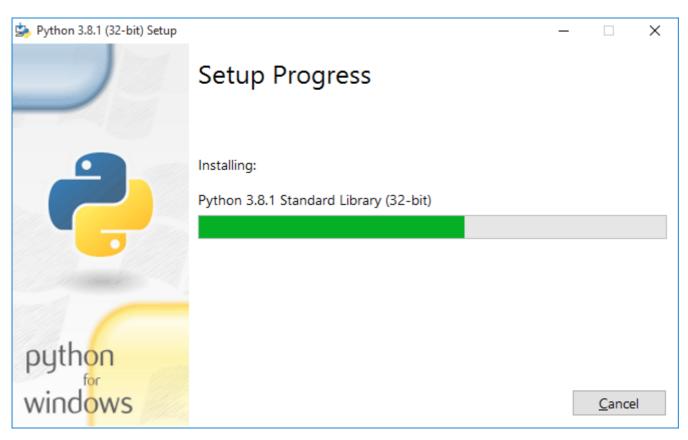
Step 4 - (Optional) Java JDK installation

If Java 8/11 is available in your system, you don't need install it again.

Install Python

Python is required for using PySpark. Follow these steps to install Python.

1) Download and install python from this web page: https://www.python.org/downloads/.



2) Verify installation by running the following command in Command Prompt or PowerShell:

```
python --version
```

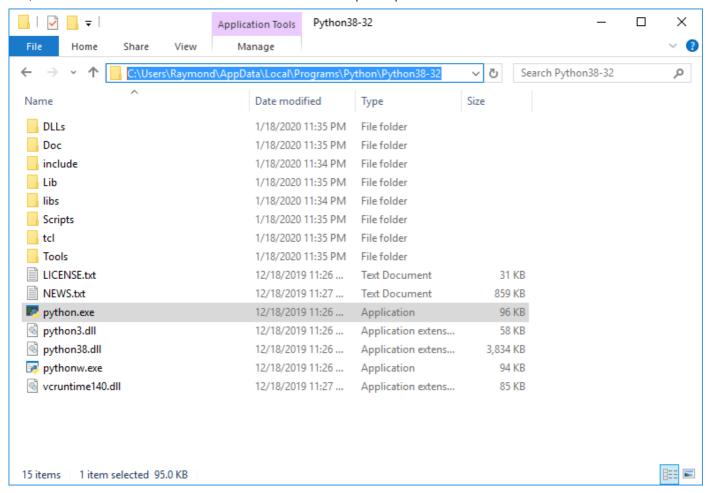
The output looks like the following:

```
Windows PowerShell
Windows PowerShell
Copyright (C) 2015 Microsoft Corporation. All rights reserved.

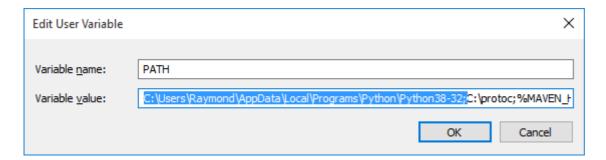
PS C:\Users\Raymond> python --version
Python 3.8.1
PS C:\Users\Raymond>
```

If python command cannot be directly invoked, please check **PATH** environment variable to make sure Python installation path is added:

For example, in my environment Python is installed at the following location:



Thus path C:\Users\Raymond\AppData\Local\Programs\Python\Python38-32 is added to PATH variable.



Hadoop installation (optional)

To work with Hadoop, you can configure a Hadoop single node cluster following this article:

Install Hadoop 3.3.0 on Windows 10 Step by Step Guide

Download binary package

Go to the following site:

https://spark.apache.org/downloads.html

Select the package type accordingly. I already have Hadoop 3.3.0 installed in my system, thus I selected the following:

Download Apache Spark™

- 1. Choose a Spark release: 3.0.0 (Jun 18 2020) ✓
- 2. Choose a package type: Pre-built with user-provided Apache Hadoop ▼
- 3. Download Spark: spark-3.0.0-bin-without-hadoop.tgz
- 4. Verify this release using the 3.0.0 signatures, checksums and project release KEYS.

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala

You can choose the package with pre-built for Hadoop 3.2 or later.

3.0.0 (Jun 18 2020) ~ Pre-built with user-provided Apache Hadoop ∨ Pre-built for Apache Hadoop 2.7 Pre-built for Apache Hadoop 3.2 and later Pre-built with user-provided Apache Hadoop Source Code

Save the latest binary to your local drive. In my case, I am saving the file to folder: F:\big-data. If you are saving the file into a different location, remember to change the path in the following steps accordingly.

Unpack binary package

Open Git Bash, and change directory (cd) to the folder where you save the binary package and then unzip using the following commands:

```
$ mkdir spark-3.0.0
$ tar -C spark-3.0.0 -xvzf spark-3.0.0-bin-without-hadoop.tgz --strip 1
```

The first command creates a sub folder named spark-3.0.0; the second command unzip the downloaded package to that folder.



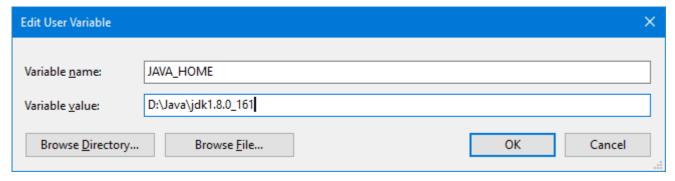
Your file name might be different from spark-3.0.0-bin-without-hadoop.tgz if you chose a package with pre-built Hadoop libs.

Spark 3.0 files are now extracted to **F:\big-data\spark-3.0.0**.

Setup environment variables

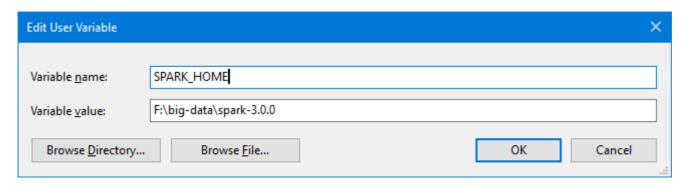
1) Setup **JAVA_HOME** variable.

Setup environment variable **JAVA_HOME** if it is not done yet. The variable value points to your Java JDK location.



2) Setup **SPARK_HOME** variable.

Setup **SPARK_HOME** environment variable with value of your spark installation directory.



3) Update **PATH** variable.

Added '%SPARK_HOME%\bin' to your **PATH** environment variable.

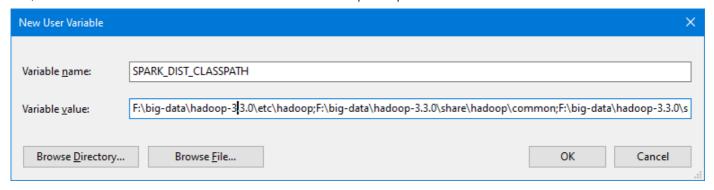


4) Configure Spark variable **SPARK_DIST_CLASSPATH**.

This is only required if you configure Spark with an existing Hadoop. If your package type already includes pre-built Hadoop libraries, you don't need to do this.

Run the following command in Command Prompt to find out existing Hadoop classpath:

Setup an environment variable SPARK_DIST_CLASSPATH accordingly using the output:



Config Spark default variables

Run the following command to create a default configuration file:

Open spark-defaults.conf file and add the following entries:

spark.driver.host localhost

Now Spark is available to use.

Verify the installation

Let's run some verification to ensure the installation is completed without errors.

Verify spark-shell command

Run the following command in Command Prompt to verify the installation.

spark-shell

The screen should be similar to the following screenshot:

```
F:\big-data\spark-shell
Setting default log level to "WARN".

Io adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel). Spark context Web UI available at http://localhost:4040

Spark context available as 'sc' (master = local[*], app id = local-1596956730299).

Spark session available as 'spark'.

Welcome to

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server UM, Java 1.8.0_161)

Type in expressions to have then evaluated.

Type:help for more information.

scala\ val site="kontext.tech"
site: String = kontext.tech"

scala\
```

You can use Scala in this interactive window.

Run examples

Execute the following command in Command Prompt to run one example provided as part of Spark installation (class SparkPi with param 10).

https://spark.apache.org/docs/latest/

%SPARK HOME%\bin\run-example.cmd SparkPi 10

The output looks like the following:

```
Select Command Prompt
2020-08-09 16:40:18,388 INFO executor.Executor: Finished task 8.0 in stage 0.0 (TID 8). 914 bytes result sent to driver
2020-08-09 16:40:18,388 INFO executor.Executor: Finished task 9.0 in stage 0.0 (TID 9). 914 bytes result sent to driver
2020-08-09 16:40:18,391 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 0.0 (TID 8) in 48 ms on raymond-pc.msh
ome.net (executor driver) (9/10)
2020-08-09 16:40:18,392 INFO scheduler.TaskSetManager: Finished task 9.0 in stage 0.0 (TID 9) in 45 ms on raymond-pc.msh
ome.net (executor driver) (10/10)
2020-08-09 16:40:18,394 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
2020-08-09 16:40:18,395 INFO scheduler.DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38) finished in 1.761 s
2020-08-09 16:40:18,404 INFO scheduler.DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks
for this job
2020-08-09 16:40:18,405 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
2020-08-09 16:40:18,407 INFO scheduler.DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 1.829346 s
Pi is roughly 3.1425391425391425
2020-08-09 16:40:18,420 INFO server.AbstractConnector: Stopped Spark@cd7f1ae{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2020-08-09 16:40:18,421 INFO ui.SparkUI: Stopped Spark web UI at http://raymond-pc.mshome.net:4040
2020-08-09 16:40:18,437 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2020-08-09 16:40:18,447 INFO memory.MemoryStore: MemoryStore cleared
2020-08-09 16:40:18,447 INFO storage.BlockManager: BlockManager stopped
2020-08-09 16:40:18,455 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
2020-08-09 16:40:18.458 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator
stopped!
2020-08-09 16:40:18,465 INFO spark.SparkContext: Successfully stopped SparkContext
2020-08-09 16:40:18,468 INFO util.ShutdownHookManager: Shutdown hook called
2020-08-09 16:40:18,469 INFO util.ShutdownHookManager: Deleting directory C:\Users\fahao.000\AppData\Local\Temp\spark-f8
35f563-9757-4d06-9695-6d52efbb6aec
2020-08-09 16:40:18,470 INFO util.ShutdownHookManager: Deleting directory C:\Users\fahao.000\AppData\Local\Temp\spark-44
2937b1-99df-4a1b-af63-11f7a0429d34
F:\big-data>
```

PySpark interactive window

Run the following command to try PySpark:

pyspark

Python in my environment is 3.8.2.

Try Spark SQL

Spark SQL interactive window can be run through this command:

```
spark-sql
```

As I have not configured Hive in my system, thus there will be error when I run the above command.

Spark context UI

When a Spark session is running, you can view the details through UI portal. As printed out in the interactive session window, Spark context Web UI available at http://localhost:4040. The URL is based on the Spark default configurations. The port number can change if the default port is used.

The following is a screenshot of the UI:

Spark Jobs (?)

User: fahao Total Uptime: 15 s Scheduling Mode: FIFO

▼ Event Timeline ○ Enable zooming

Enable 200ming											
Executors											
Added											
Removed											
Jobs											
Succeeded											
Failed											
Running											
	200	400	600	800	000	200	400	600	800	000	200
	17:05:28				17:05:29					17:05:30	

References

Spark developer tools

Refer to the following page if you are interested in any Spark developer tools.

https://spark.apache.org/developer-tools.html

Spark 3.0.0 overview

Refer to the official documentation about Spark 3.0.0 overview: http://spark.apache.org/docs/3.0.0/.

Spark 3.0.0 release notes

https://spark.apache.org/releases/spark-release-3-0-0.html

Congratulations! You have successfully configured Spark in your Windows environment. Have fun with Spark 3.0.0.

More from Kontext

- Apache Hive 3.0.0 Installation on Windows 10 Step by Step Guide
- Apache Hive 3.1.1 Installation on Windows 10 using Windows Subsystem for Linux
- Apache Spark 3.0.0 Installation on Linux Guide

- Apache Spark 2.4.3 Installation on Windows 10 using Windows Subsystem for Linux
- Install Apache Sqoop in Windows
- Install Hadoop 3.0.0 on Windows (Single Node)

spark

pyspark

windows10

big-data-on-windows-10

(i) Last modified by Raymond 3 years ago

© This page is subject to Site terms.

Like this article?



Share on





Comments



Raymond

(2 years ago 😄 🚦



#1528 Re: Install Apache Spark 3.0.0 on Windows 10

If you use Derby for hive metastore, please ensure that the directory context in your command prompt is the same when you run your previous init command previously otherwise you will have to initialize the metastore again. I feel like the error you got was caused by that but I will need to look into details to be able to tell.

For the data warehouse folder, it exists in HDFS not in file system directly.



Orland () 2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

Nope it didnt. BTw Raymond I managed to run my hive smoothly the other day after installation and was able to access the hiveserver2 but now when I try to connect Im able to access hive but the hive --help doesnt work and I cant connect to the hiveserver2 as well when I run these commands:

HIVE_HOME/bin/hive --service metastore &

\$HIVE_HOME/bin/hive --service hiveserver2 start &

also I dont have hive in my users directory with a warehouse subfolder /user/hive/warehouse.





(1) 2 years ago 😄 :

#1525 Re: Install Apache Spark 3.0.0 on Windows 10

Nope it didnt. BTw Raymond I managed to run my hive smoothly the other day after installation and was able to access the hiveserver2 but now when I try to connect Im able to access hive but the hive --help doesnt work and I cant connect to the hiveserver2 as well when I run these commands:

HIVE HOME/bin/hive --service metastore &

\$HIVE HOME/bin/hive --service hiveserver2 start &

also I dont have hive in my users directory with a warehouse subfolder /user/hive/warehouse.



2 Raymond () 2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

Did your Spark session crash after you see the warning message?

WARN executor.ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped.

If it doesn't crash, it is ok. There was recommendation of creating PYSPARK_PYTHON ...

+ Read more





Raymond

() 2 years ago 😝 :

#1523 Re: Install Apache Spark 3.0.0 on Windows 10

Did your Spark session crash after you see the warning message?

WARN executor.ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped.

If it doesn't crash, it is ok. There was recommendation of creating PYSPARK_PYTHON ...

+ Read more



© Orland © 2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

This is the warning. Thank you.

□ Command Propet - spak shall

Wicrosoft Windows (Version 10.8.19843.1227)
(c) Microsoft Windows (Version 10.8.19843.1227)
(c) Microsoft Corporation. All rights reserved.

C. Users/Users/Spark. Shell

Setting default log level to NARP.
To adjust logging level use sc. settoglevel/(newLevel). For SparkR, use settoglevel/(newLevel).

Spark context available as 'sc ([master = local*]), app id = local*1634481462874).

Welcome to

Using Scala version 2.13.18 (Java MotSpet(TM) 64-Bit Server VM, Java 1.8.0.381)
Type in expressions to have these evaluated.

Type in expressions to have the evaluated.

Type in expressions to have these evaluated.





(2 years ago 🖨 :

#1522 Re: Install Apache Spark 3.0.0 on Windows 10

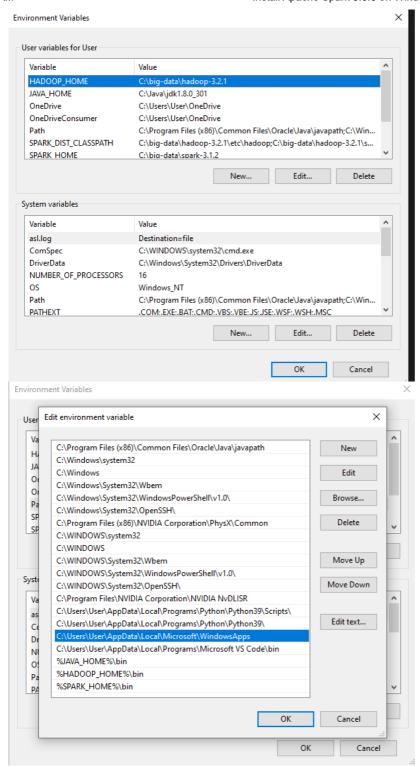
Sorry I forgot to mention that **%SPARK_HOME**% works with Command Prompt. For Git Bash, please use **\$SPARK_HOME** to access the environment variable:



For adding Git Bash bin to **PATH** variable: please add path **C:\Program Files\Git**\usr\bin to environment variable **PATH**. Depends on where Git is installed in your computer, please change the path accordingly. You can just directly go to Spark installation folder and then manually copy the file instead of using command.



Orland © 2 years ago
Re: Install Apache Spark 3.0.0 on Windows 10
How do I setup git bash in path?







0 Orland (2 years ago 😄



#1521 Re: Install Apache Spark 3.0.0 on Windows 10

This is the warning. Thank you.

ll
dito "MARN".
use sc.setloglevel(newLevel). For SparkR, use setloglevel(newLevel).
silable at http://localhost:4040
as 'sc' (master = local[*], app id = local-1634481462874). 12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_301) have them evaluated. -10-17 22:38:00,498 WARN executor.ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result of ProcessTree metrics is stopped



2 Raymond () 2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

Hi Orland,

For copying file, have your opened a new window after setting the environment variable SPARK_HOME. For terminal opened before you set the variable, it won't be effective. Also cp command only exists in PowerShell or Git Bash or Command Prompt (when you have added Git Bash bin folder to the PATH).

For the error you got, it is actually a warning message and I think you can just ignore it. Let me know if your whole process cannot wrong because of that error. BTW, it will be helpful if you provide screenshot so that I can view all the error messages.

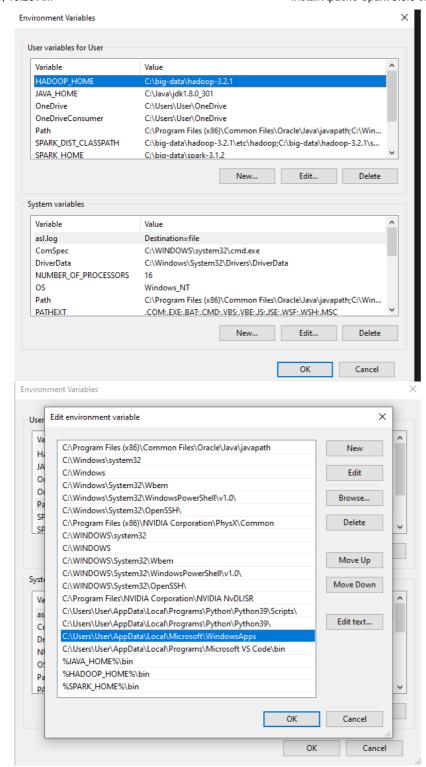
Reply

O Orland

() 2 years ago 😝 :

#1520 Re: Install Apache Spark 3.0.0 on Windows 10

How do I setup git bash in path?



```
### MINGW64:/c/Users/User

### efaults.conf'

### cp: cannot stat '%SPARK_HOME%/conf/spark-defaults.conf.template': No such file in directory

### directory
```

99

2 Raymond (2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

Hi Orland,

For copying file, have your opened a new window after setting the environment variable SPARK_HOME. For terminal opened before you set the variable, it won't be effective. Also cp command only exists in PowerShell or Git Bash or Command Prompt (when you have added Git Bash bin folder to the PATH).

For the error you got, it is actually a warning message and I think you can just ignore it. Let me know if your whole process cannot wrong because of that error. BTW, it will be helpful if you provide screenshot so that I can view all the error messages.







#1519 Re: Install Apache Spark 3.0.0 on Windows 10

Hi Orland,

For copying file, have your opened a new window after setting the environment variable SPARK_HOME. For terminal opened before you set the variable, it won't be effective. Also cp command only exists in PowerShell or Git Bash or Command Prompt (when you have added Git Bash bin folder to the PATH).

For the error you got, it is actually a warning message and I think you can just ignore it. Let me know if your whole process cannot wrong because of that error. BTW, it will be helpful if you provide screenshot so that I can view all the error messages.



Orland () 2 years ago

Re: Install Apache Spark 3.0.0 on Windows 10

Hi Raymond Im getting this error when I run Spark and pyspark in command prompt. How to fix it?

WARN executor. Procfs Metrics Getter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped.

I also tried typing in 'cp %SPARK_HOME%/conf/spark-defaults.conf.template %SPARK_HOME%/conf/spark-defaults.conf' in command prompt and git bash but it wasnt recognized.

Thank you.





(C) 2 years ago 😝 🚦



#1518 Re: Install Apache Spark 3.0.0 on Windows 10

Hi Raymond Im getting this error when I run Spark and pyspark in command prompt. How to fix it?

WARN executor. Procfs Metrics Getter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped.

I also tried typing in 'cp %SPARK_HOME%/conf/spark-defaults.conf.template %SPARK_HOME%/conf/spark-defaults.conf' in command prompt and git bash but it wasnt recognized.

Thank you.



Please log in or register to comment.

② Log in +2 Register

Log in with external accounts

Log in with Microsoft account

G Log in with Google account

i Table of contents

i Stats

Big Data Tools on Windows 10

spark

pyspark

windows10

Home / Columns / Spark & PySpark / Install Apache Spark 3.0.0 on Windows 10

Kontext

© Kontext 2023 v1.2.7

Made with ♥ in Melbourne.



Products & features

Columns

Forums

Tags

Series

Search

Resources

Subscribe RSS

Create your Column

Help centre

Cookie

Privacy

Terms

Contact us

Subscription Subscribe to Kontext newsletter to get updates about data analytics, programming and cloud related articles.

Email

➤ Subscribe