Kontext

Install Apache Spark 3.0.0 on Windows 10

Raymond 9,782 8 2020-08-09 3 years ago

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.

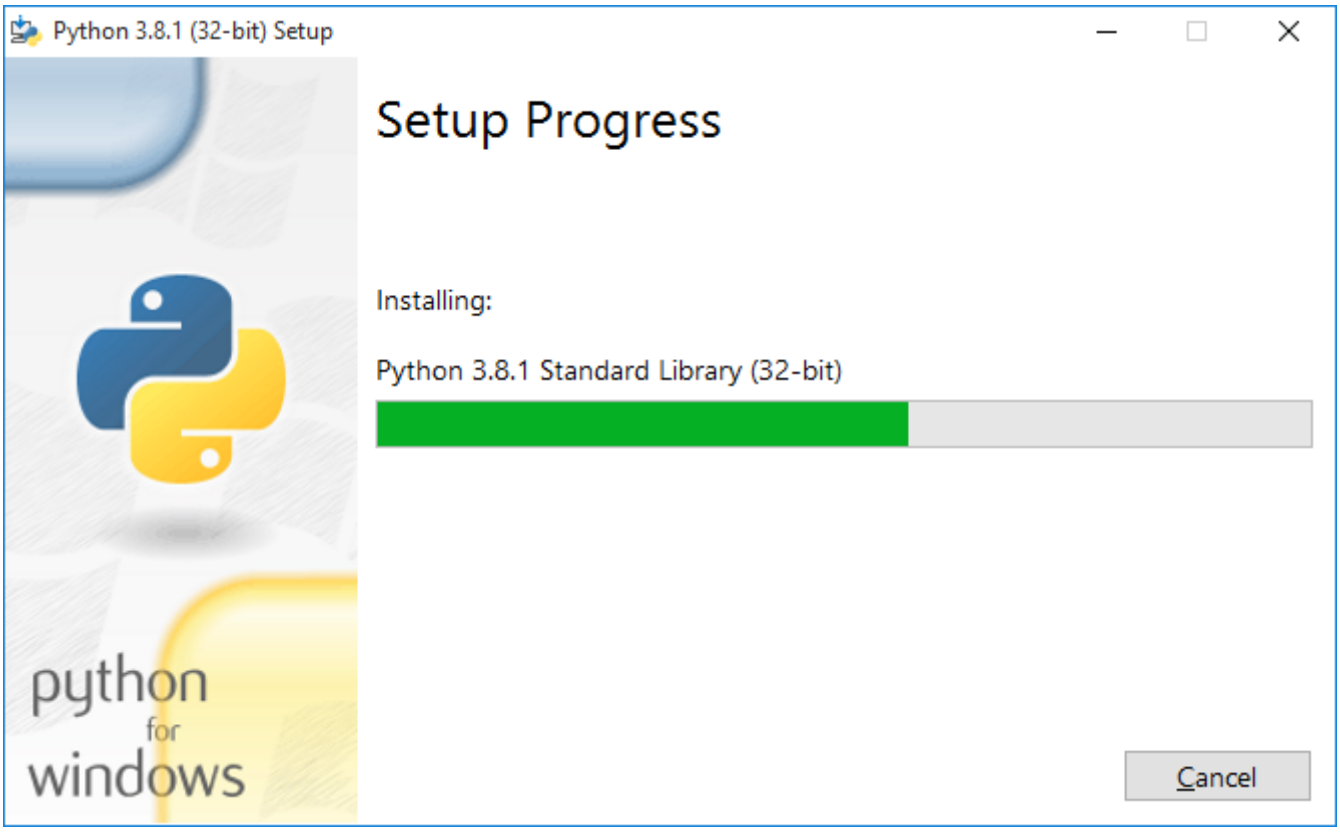
Step 4 - (Optional) Java JDK installation

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.

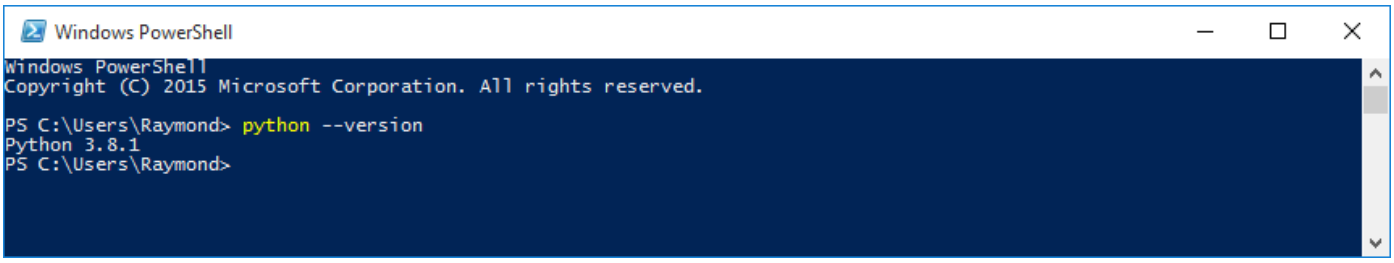
1) Download and install python from this web page: https://www.python.org/downloads/.



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

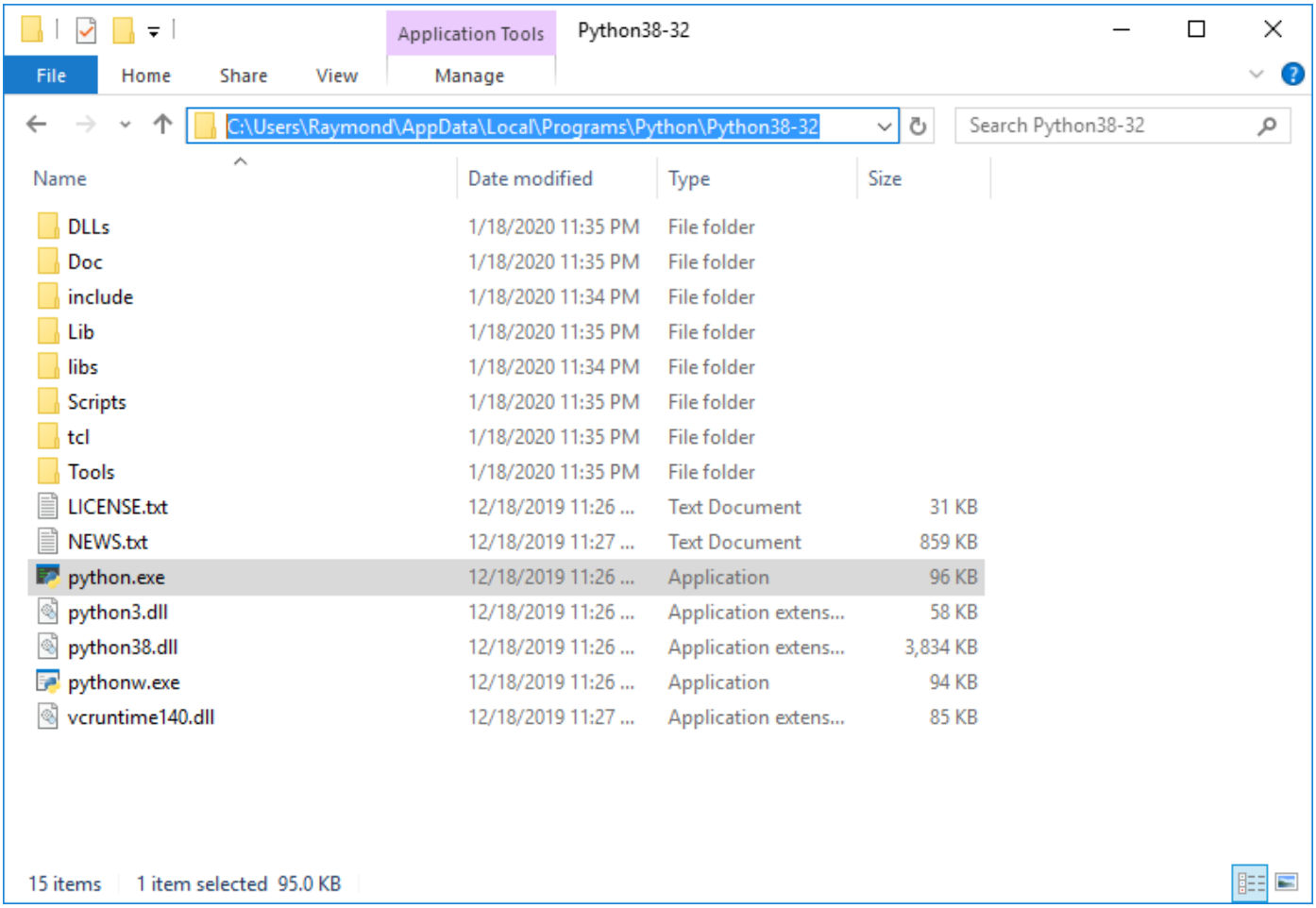
```
python --version
```

The output looks like the following:

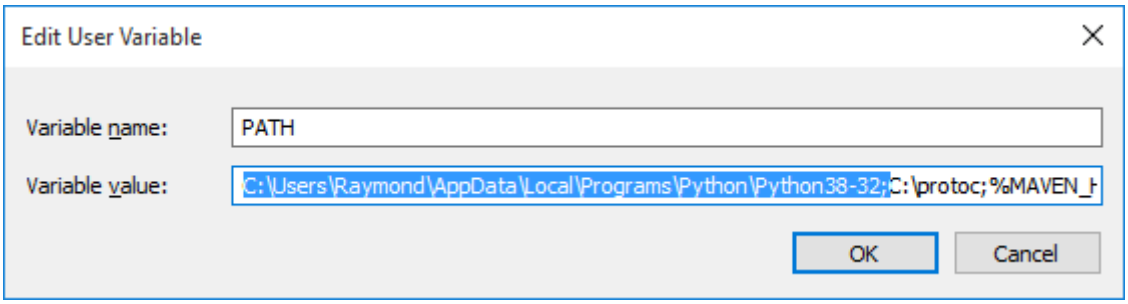


If python command cannot be directly invoked, please check **PATH** environment variable to make sure Python installation path is added:

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:

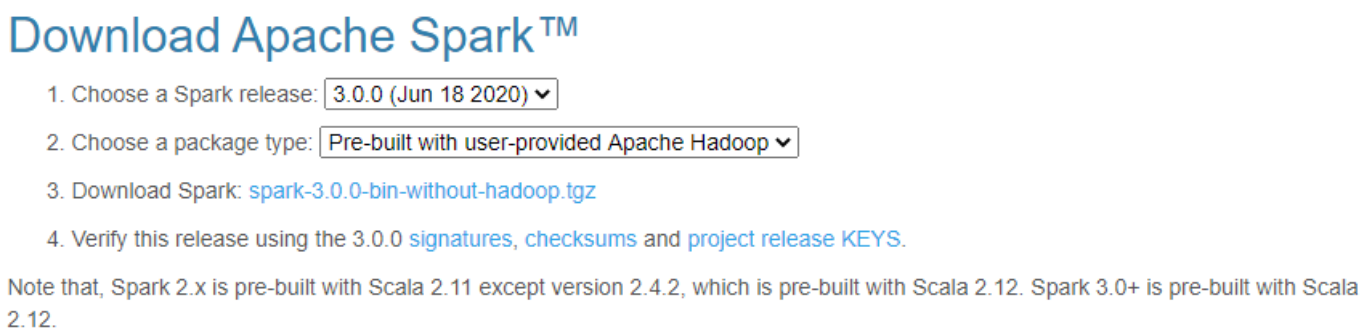
Install Hadoop 3.3.0 on Windows 10 Step by Step Guide

Download binary package

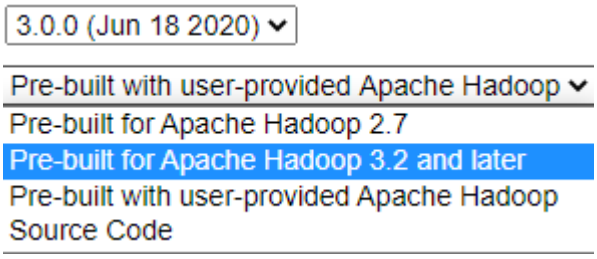
Go to the following site:

> 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:



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



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.

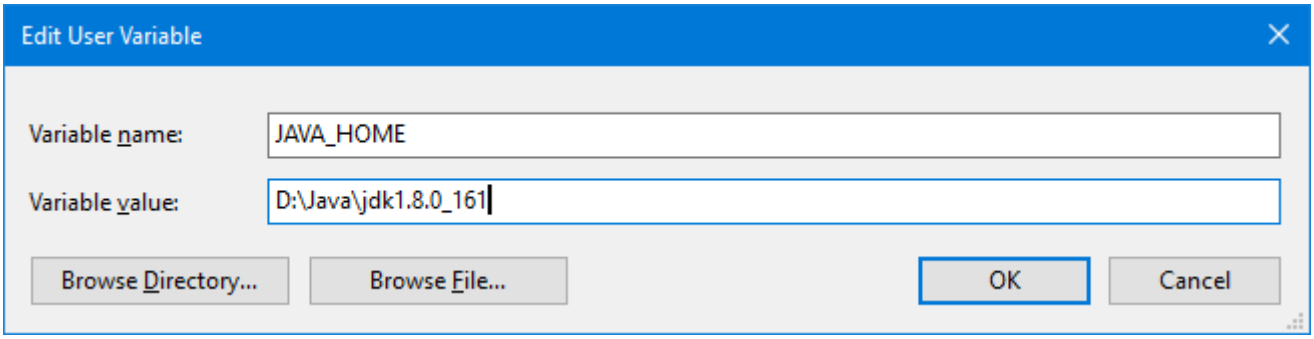
> ⚠ Your file name might be different from **spark-3.0.0-bin-without-hadoop.tgz** if you chose a package with pre-built Hadoop libs.

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

Setup environment variables

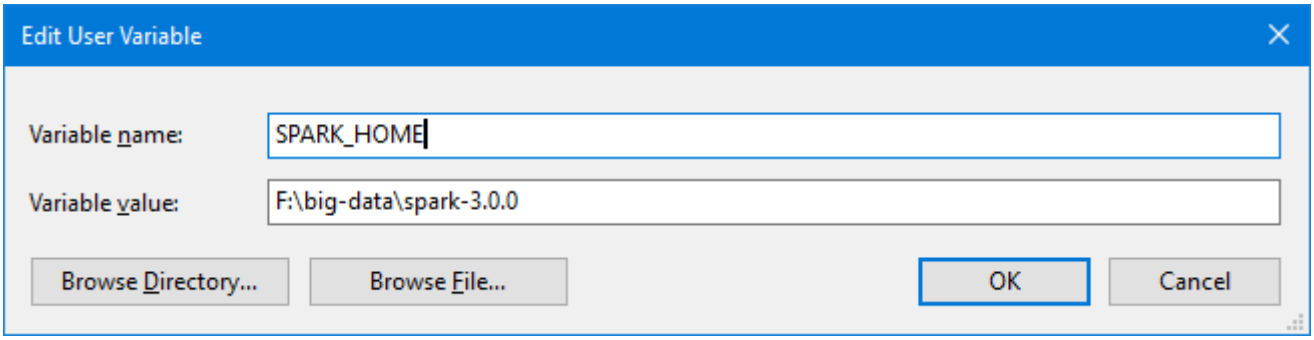
1) Setup **JAVA_HOME** variable.

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



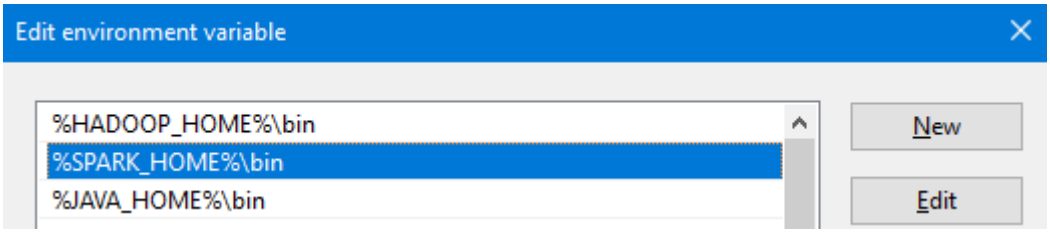
2) Setup **SPARK_HOME** variable.

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.



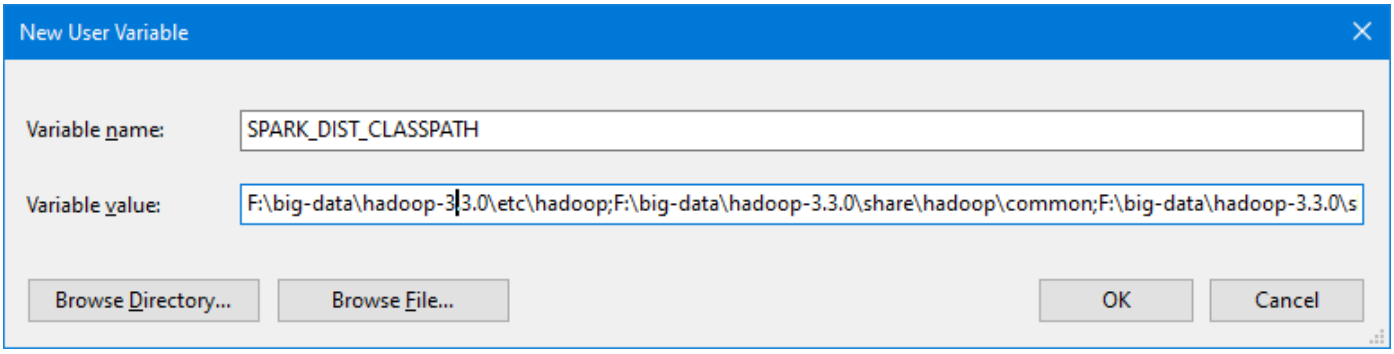
4) Configure Spark variable **SPARK_DIST_CLASSPATH**.

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:

```
F:\big-data>hadoop classpath
F:\big-data\hadoop-3.3.0\etc\hadoop;F:\big-data\hadoop-
3.3.0\share\hadoop\common;F:\big-data\hadoop-3.3.0\share\hadoop\common\lib\*;F:\big-
data\hadoop-3.3.0\share\hadoop\common\*;F:\big-data\hadoop-
3.3.0\share\hadoop\hdfs;F:\big-data\hadoop-3.3.0\share\hadoop\hdfs\lib\*;F:\big-
data\hadoop-3.3.0\share\hadoop\hdfs\*;F:\big-data\hadoop-
3.3.0\share\hadoop\yarn;F:\big-data\hadoop-3.3.0\share\hadoop\yarn\lib\*;F:\big-
data\hadoop-3.3.0\share\hadoop\yarn\*;F:\big-data\hadoop-3.3.0\share\hadoop\mapreduce\*
```

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:

```
cp %SPARK_HOME%/conf/spark-defaults.conf.template %SPARK_HOME%/conf/spark-defaults.conf
```

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

```
spark.driver.host       localhost
```

Now Spark is available to use.

Verify the installation

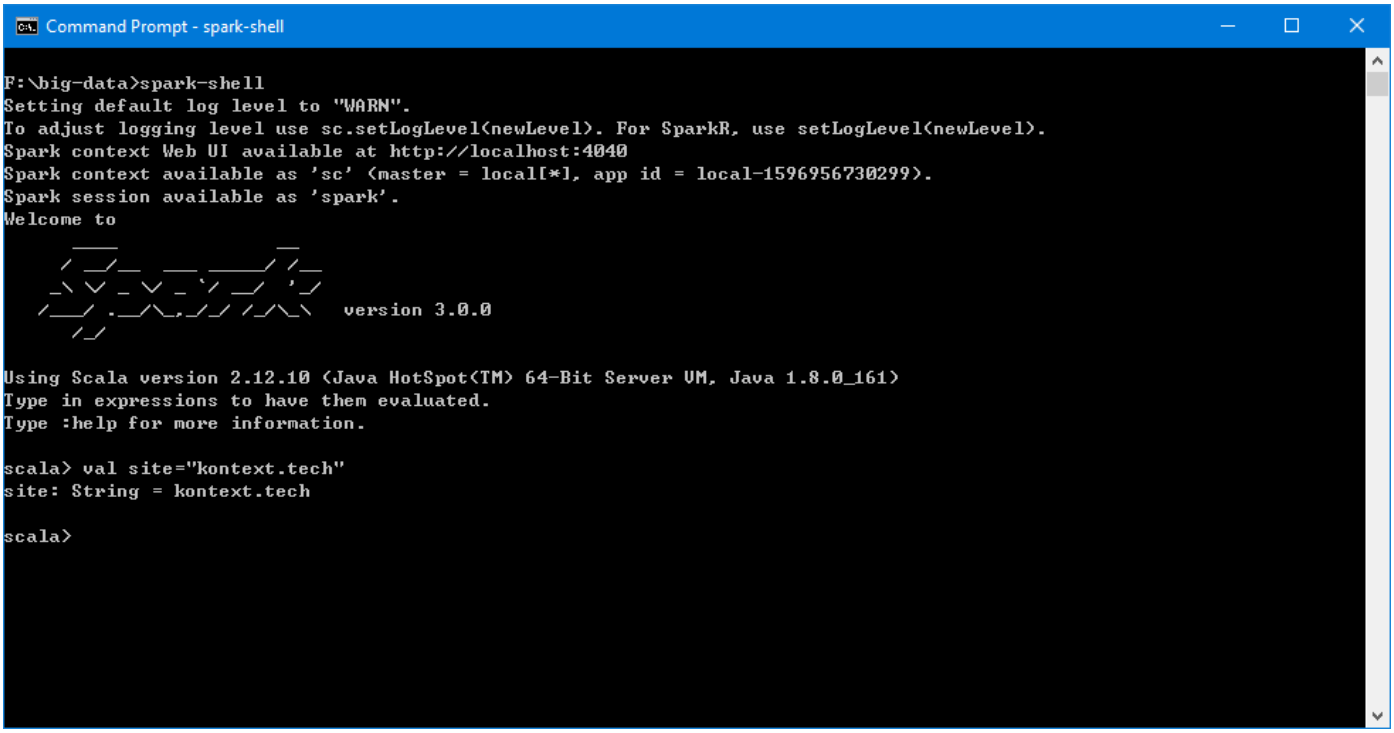
Let's run some verification to ensure the installation is completed without errors.

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:



You can use Scala in this interactive window.

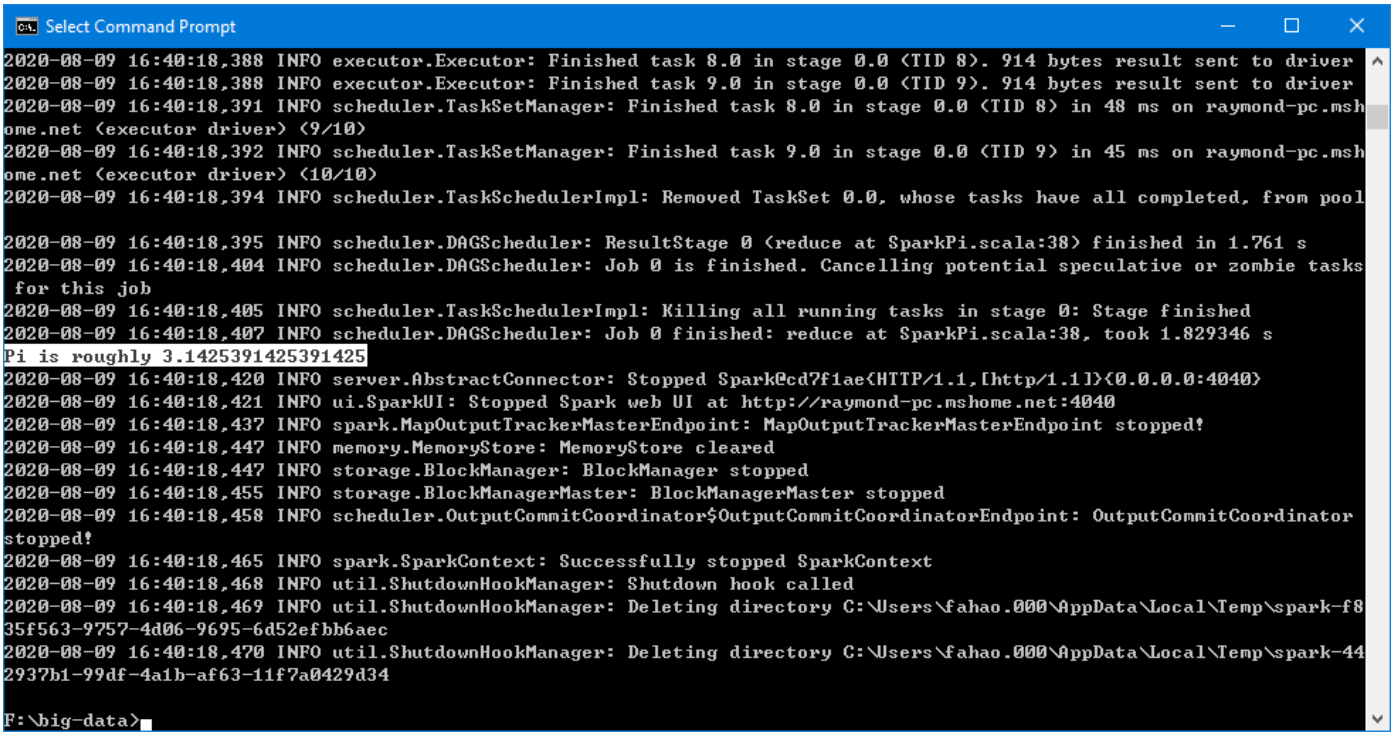
Run examples

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

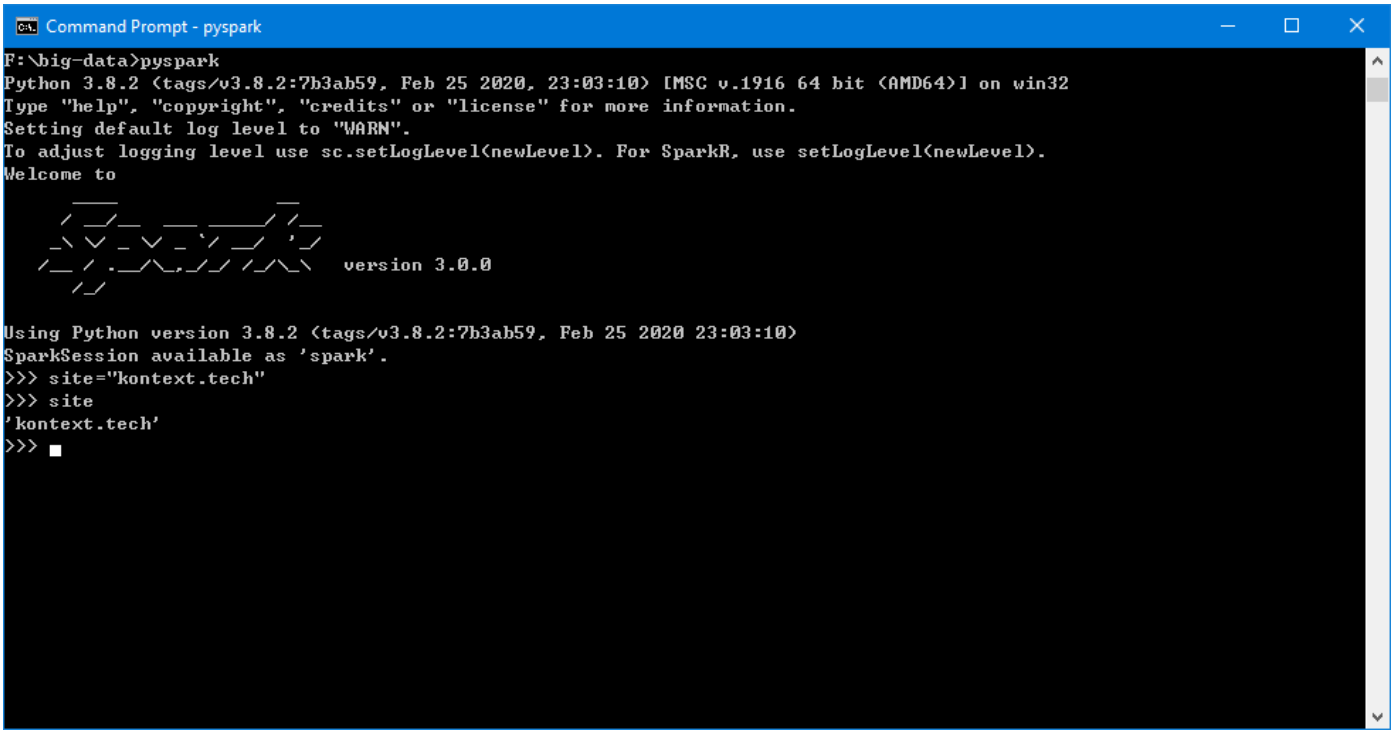
The output looks like the following:



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:

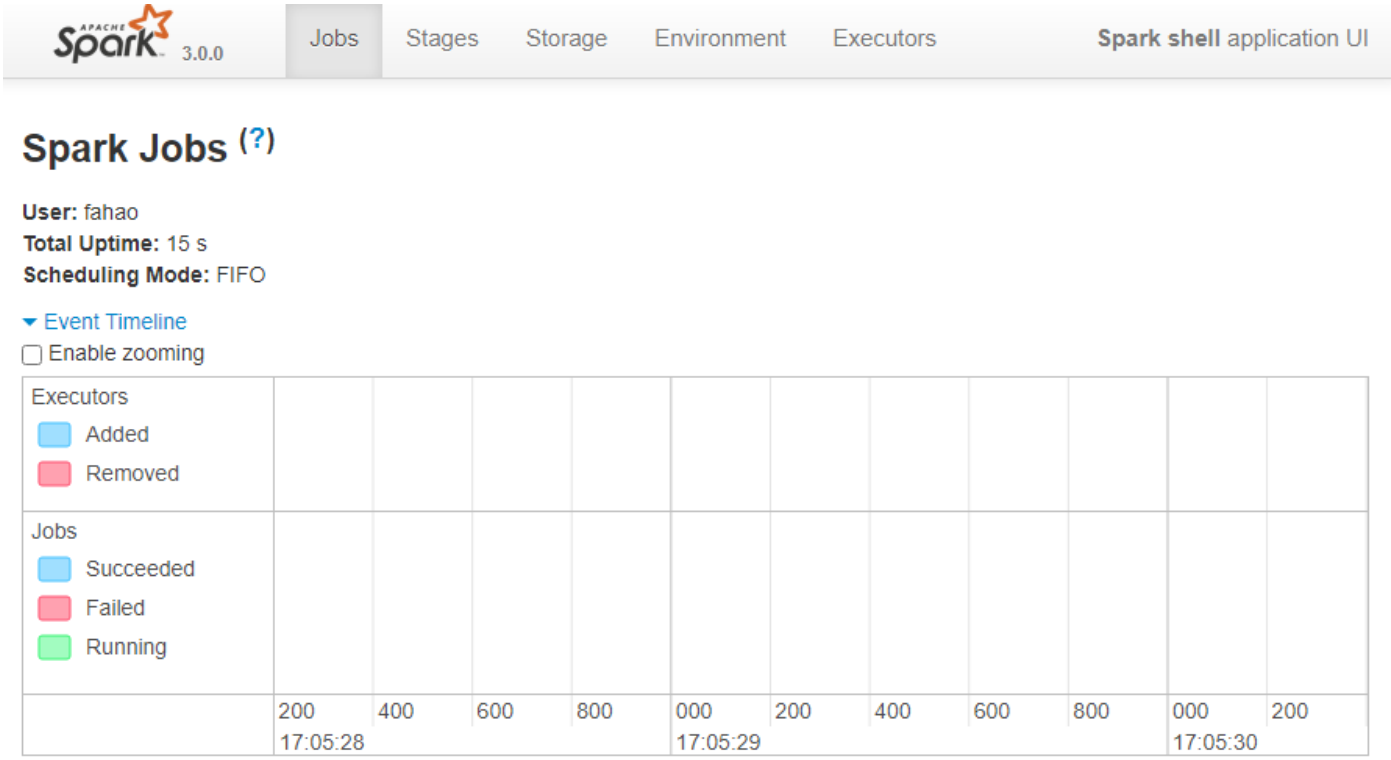
```
spark-sql
```

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:



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

Last modified by Raymond 3 years ago

Like this article?

2

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.

Reply

O Orland

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.

> 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

Reply

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.
>
> 
>

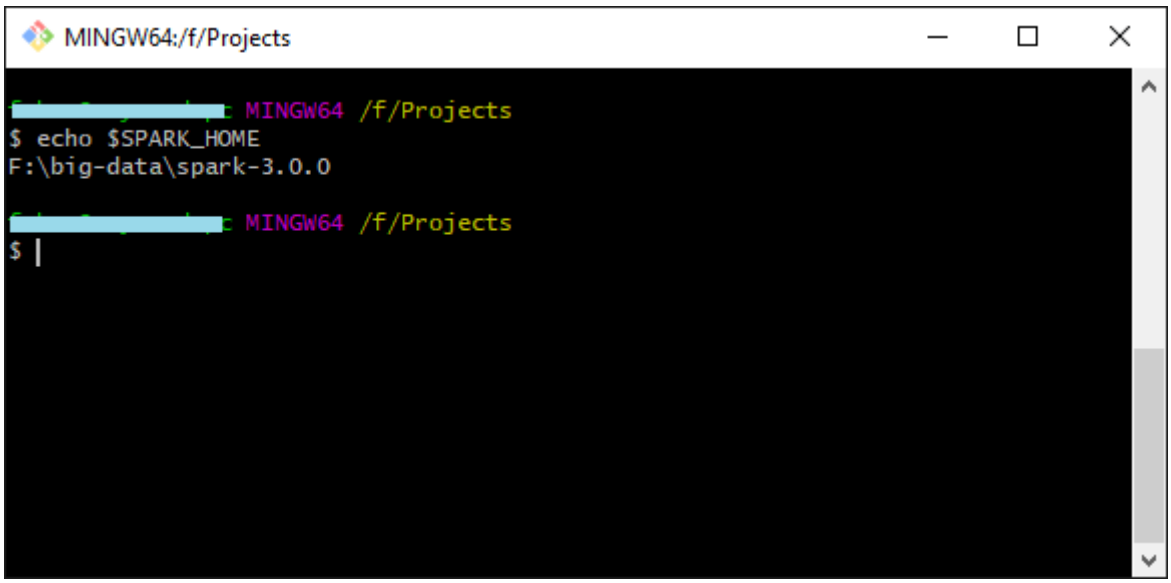
Reply

Raymond

2 years ago

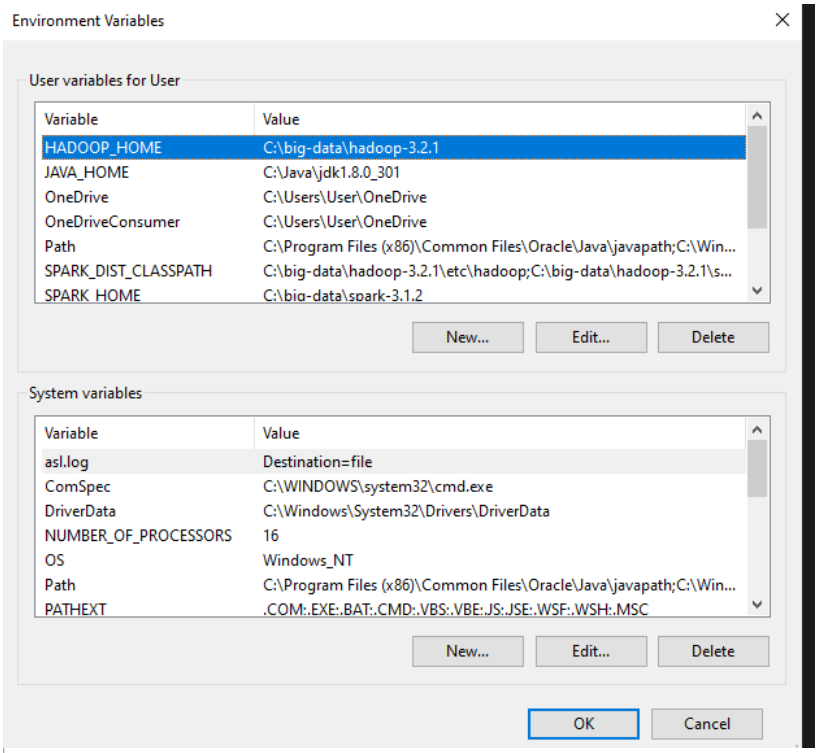
#1522 Re: Install Apache Spark 3.0.0 on Windows 10

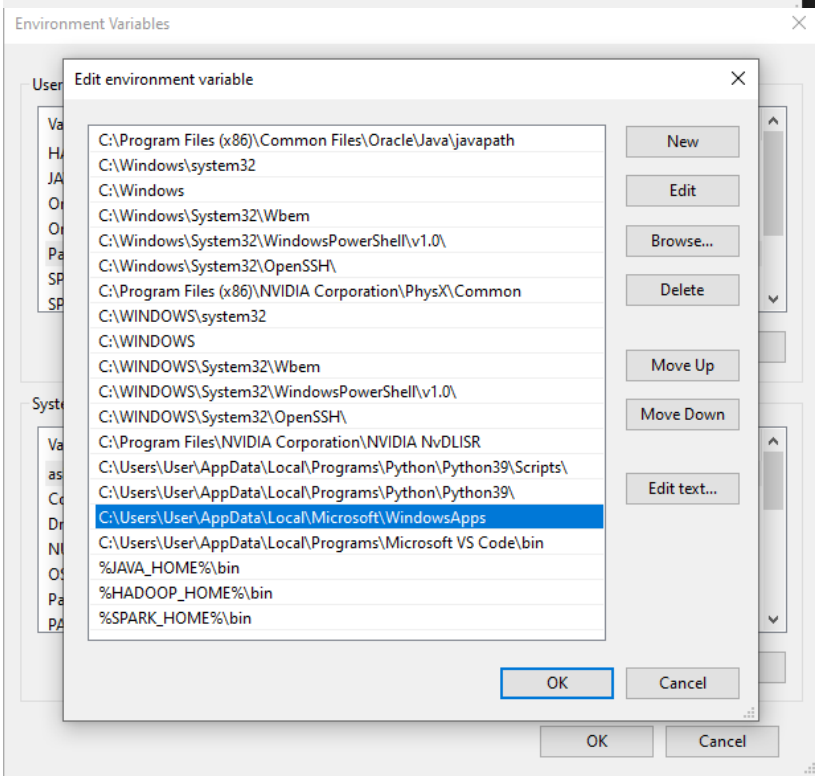
Sorry I forgot to mention that **%SPARK_HOME%** works with Command Prompt.
For Git Bash, please use **$SPARK_HOME** to access the environment variable:

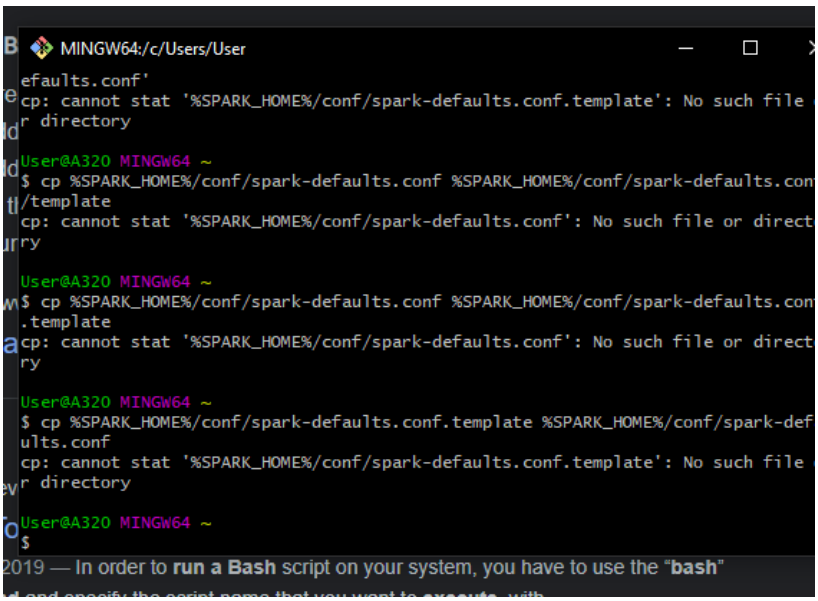


For adding Git Bash bin to **PATH** variable: please add path **C:\Program Files\Git**\usr\bin to environment variable **PATH**. Depends on where Git is installed in your computer, please change the path accordingly. You can just directly go to Spark installation folder and then manually copy the file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


Environment Variables
User variables for User
Variable
Value
HADOOP_HOME
C:\big-data\hadoop-3.2.1
JAVA_HOME
C:\Java\jdk1.8.0_301
OneDrive
C:\Users\User\OneDrive
OneDriveConsumer
C:\Users\User\OneDrive
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
SPARK_DIST_CLASSPATH
C:\big-data\hadoop-3.2.1\etc\hadoop;C:\big-data\hadoop-3.2.1\s...
SPARK_HOME
C:\big-data\spark-3.1.2
New...
Edit...
Delete
System variables
Variable
Value
asl.log
Destination=file
ComSpec
C:\WINDOWS\system32\cmd.exe
DriverData
C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS
16
OS
Windows_NT
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
New...
Edit...
Delete
OK
Cancel


Environment Variables
Edit environment variable
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Windows\System32\OpenSSH\
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
C:\WINDOWS\system32
C:\WINDOWS
C:\WINDOWS\System32\Wbem
C:\WINDOWS\System32\WindowsPowerShell\v1.0\
C:\WINDOWS\System32\OpenSSH\
C:\Program Files\NVIDIA Corporation\NVIDIA NvDLISR
C:\Users\User\AppData\Local\Programs\Python\Python39\Scripts\
C:\Users\User\AppData\Local\Programs\Python\Python39\
C:\Users\User\AppData\Local\Microsoft\WindowsApps
C:\Users\User\AppData\Local\Programs\Microsoft VS Code\bin
%JAVA_HOME%\bin
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel
OK
Cancel



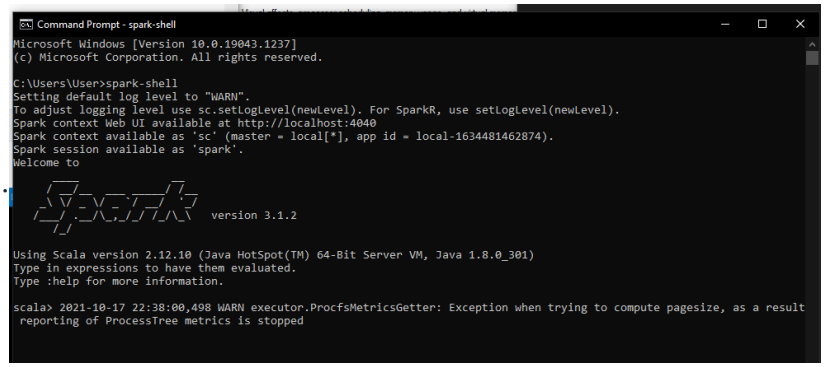
Reply

O Orland

2 years ago

#1521 Re: Install Apache Spark 3.0.0 on Windows 10

This is the warning. Thank you.



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.

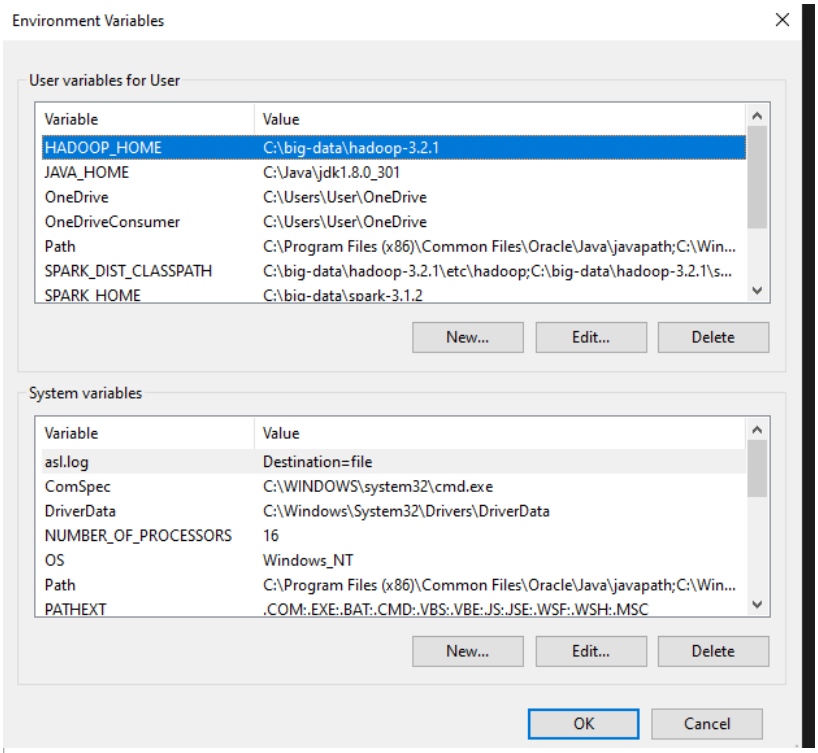
Reply

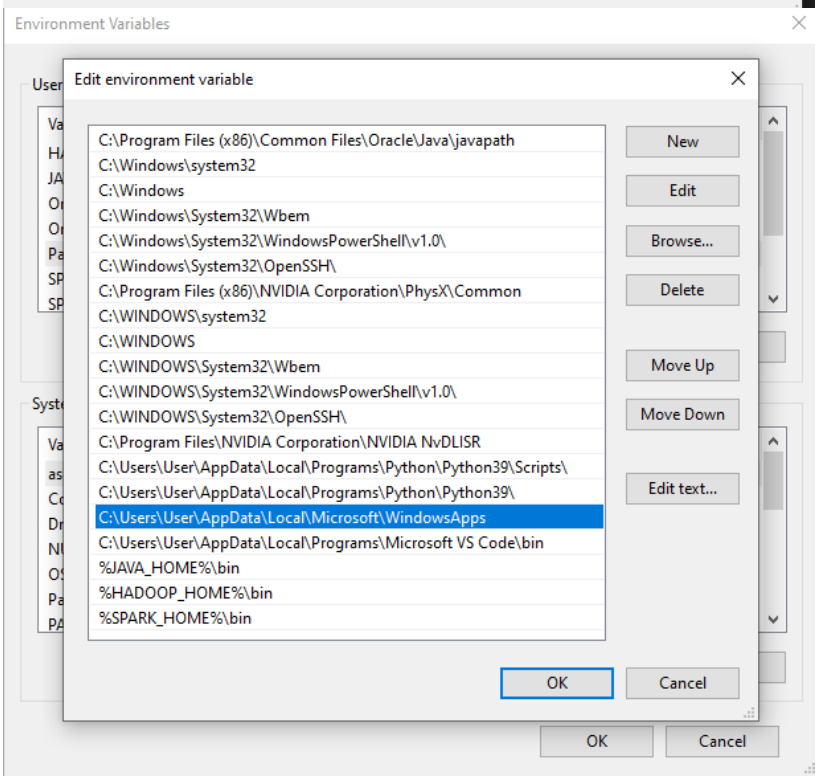
O Orland

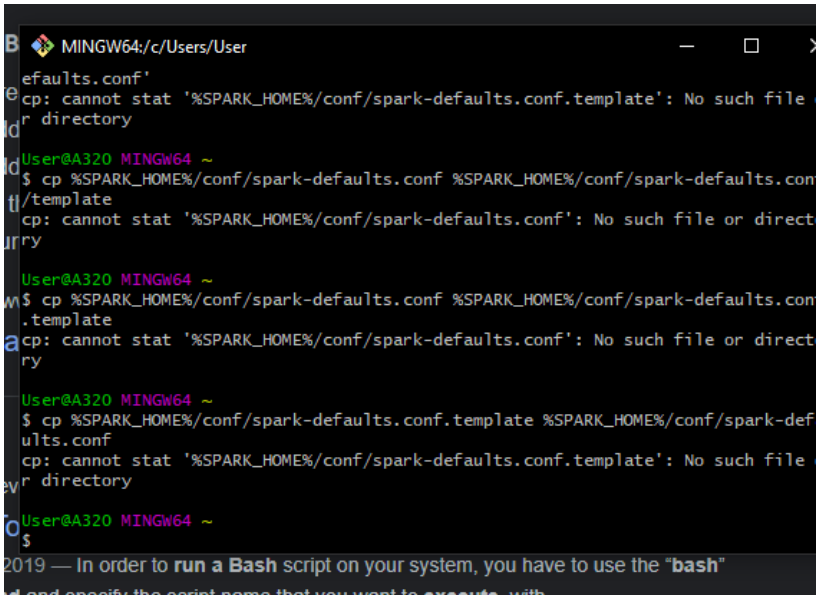
2 years ago

#1520 Re: Install Apache Spark 3.0.0 on Windows 10

How do I setup git bash in path?


Environment Variables
User variables for User
Variable
Value
HADOOP_HOME
C:\big-data\hadoop-3.2.1
JAVA_HOME
C:\Java\jdk1.8.0_301
OneDrive
C:\Users\User\OneDrive
OneDriveConsumer
C:\Users\User\OneDrive
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
SPARK_DIST_CLASSPATH
C:\big-data\hadoop-3.2.1\etc\hadoop;C:\big-data\hadoop-3.2.1\s...
SPARK_HOME
C:\big-data\spark-3.1.2
New...
Edit...
Delete
System variables
Variable
Value
asl.log
Destination=file
ComSpec
C:\WINDOWS\system32\cmd.exe
DriverData
C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS
16
OS
Windows_NT
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
New...
Edit...
Delete
OK
Cancel


Environment Variables
Edit environment variable
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Windows\System32\OpenSSH\
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
C:\WINDOWS\system32
C:\WINDOWS
C:\WINDOWS\System32\Wbem
C:\WINDOWS\System32\WindowsPowerShell\v1.0\
C:\WINDOWS\System32\OpenSSH\
C:\Program Files\NVIDIA Corporation\NVIDIA NvDLISR
C:\Users\User\AppData\Local\Programs\Python\Python39\Scripts\
C:\Users\User\AppData\Local\Programs\Python\Python39\
C:\Users\User\AppData\Local\Microsoft\WindowsApps
C:\Users\User\AppData\Local\Programs\Microsoft VS Code\bin
%JAVA_HOME%\bin
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel
OK
Cancel



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

Raymond 2 years ago

#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.ProcfsMetricsGetter: 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.

Reply

O Orland

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.ProcfsMetricsGetter: 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.

Reply

Please log in or register to comment.

Log in | Register

Log in with external accounts

Log in with Microsoft account

Log in with Google account

Table of contents

Stats

Big Data Tools on Windows 10

spark

pyspark

windows10

About

Cookie

Privacy

Terms

Contact us

Subscription

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

Email

Subscribe