

**Assignment 2**

**SEMESTER : 3 - 2017/2018**

**COURSE CODE : MANB1163**

**COURSE : CLOUD COMPUTING FOR BIG DATA ANALYTICS**

**PROGRAMME : MSc (BIA)**

**SUBMISSION DATE : 27 OCTOBER 2018**

**PREPARED BY : MUHAMMAD NIDZAM BIN MASO’OD MAN171043**

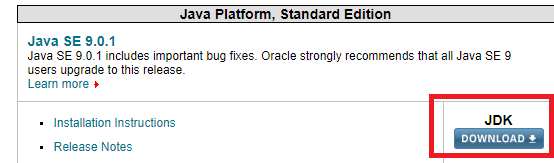
**SUBMITTED TO : DR. MOHD SYAHID BIN MOHD ANUAR**

Apache Spark and Pyspark installation manual

Contents

[1. Installing pyspark 3](#_Toc529774670)

[2. Installing Apache Spark 4](#_Toc529774671)

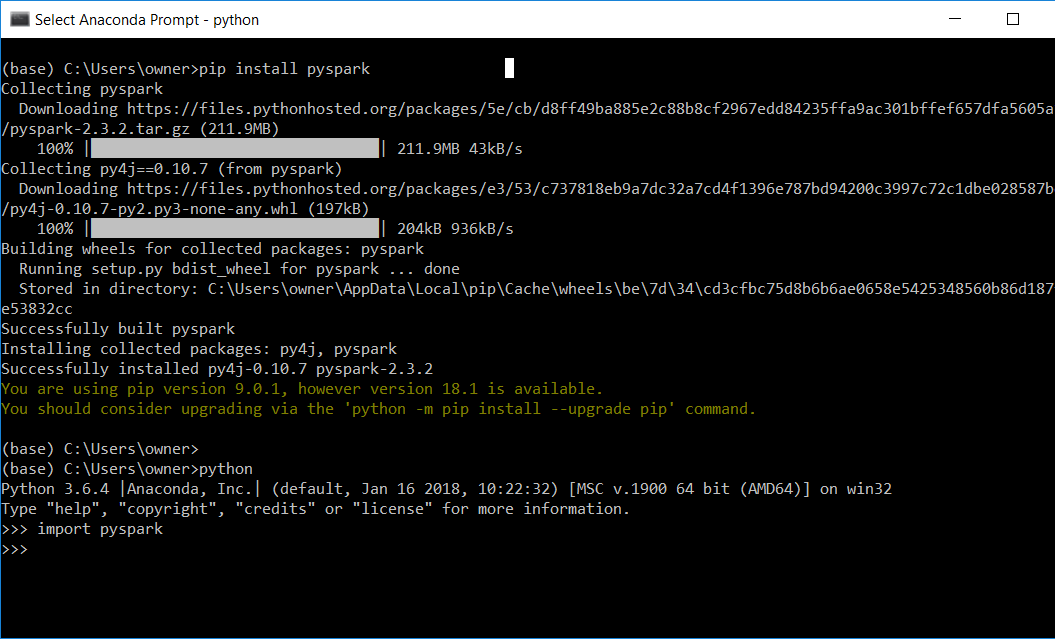
[ 5](#_Toc529774672)

# Installing pyspark

1. Start Anaconda Navigator and key in below command into command prompt

pip install pyspark

\*after installation finished, you can test the installation by importing pyspark



.

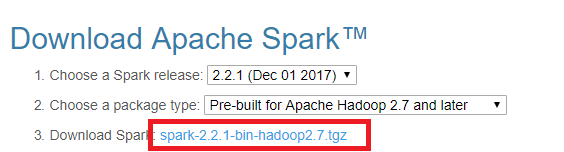
# Installing Apache Spark

1. Software preparation

To Install Apache Spark on Window 10 64-bit, the first steps are to download and install below software.

1. Spark distribution from below link

[spark.apache.org](https://spark.apache.org/downloads.html)



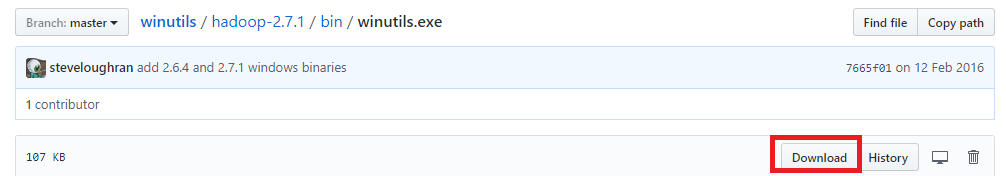
1. Python and Jupyter Notebook. You can get both by installing the Python 3.x version of Anaconda distribution.

<https://www.anaconda.com/download/>

1. Install winutils.exe — a Hadoop binary for Windows

<https://github.com/steveloughran/winutils/>

Go to the corresponding Hadoop version in the Spark distribution and find winutils.exe under /bin. For example, <https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>



1. Download and install latest Java JDK (above version 7) from below link

<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

## 

1. Spark installation
2. Unpacked .tgz file into spark folder using WINRAR

C:\spark\spark

1. Move the winutils.exe downloaded from step A-c to the \bin folder of Spark distribution

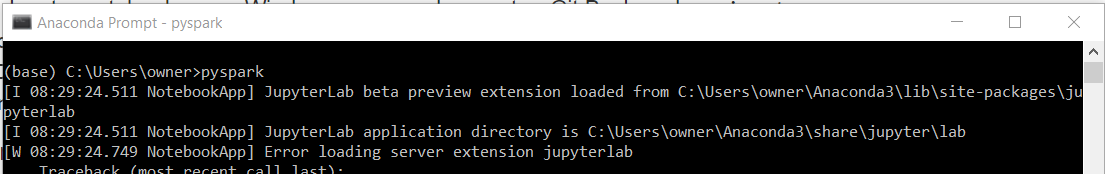
C:\spark\spark\bin

1. add environment variables: the environment variables let Windows find where the files are when we start the PySpark kernel. You can find the environment variable settings by putting “environ…” in the window search box. For my example

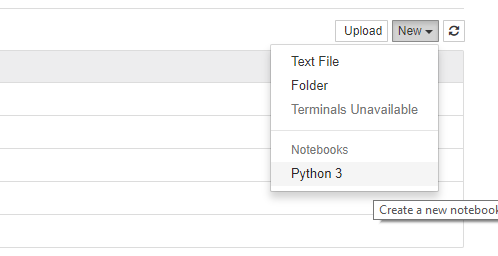
| **Name** | **Value** |
| --- | --- |
| SPARK\_HOME | D:\spark\spark-2.2.1-bin-hadoop2.7 |
| HADOOP\_HOME | C:\spark\spark |
| PYSPARK\_DRIVER\_PYTHON | jupyter |
| PYSPARK\_DRIVER\_PYTHON\_OPTS | notebook |
| JAVA\_HOME | C:\Java\jdk |

1. Running PySpark in Jupyter Notebook
2. Open ANACONDA prompt. Run the Hadoop server first. To run SPARK command enter:pyspark

(base) C:\Users\owner>pyspark



1. Once pyspark is running, it will open open jupyter notebook automatically. Create a new python 3 notebook.



1. In the notebook, run the following code

import findspark

findspark**.**init()

import pyspark *# only run after findspark.init()*

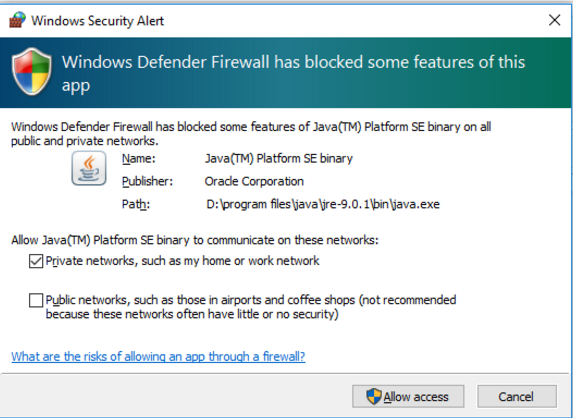
from pyspark.sql import SparkSession

spark **=** SparkSession**.**builder**.**getOrCreate()

df **=** spark**.**sql('''select 'spark' as hello ''')

df**.**show()

When you press run, it might trigger a Windows firewall pop-up. Press cancel on the pop-up as blocking the connection doesn’t affect PySpark.



1. If you see the following output, then you have installed PySpark successfully on your Windows system.

