Omgamganapathayenamaha

Omgamganapathayenamaha

Omgamganapathayenamaha

Omgamganapathayenamaha

**Installing Python via Enthought Canopy**

1. Go to <https://assets.enthought.com/downloads/>
2. Download Windows-64 bit 3.5 supported Standard Insaller
3. Click the .exe and install it
4. Right click “Enthought Canopy” icon on the Desktop and click properties
   1. In the “Compatibility” Tab, Mark “Run this Program as an Administrator”

**Installing JDK 8**

1. Apache Spark have support only with JDK 8 as of now
2. Go to <https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>
3. Accept the License agreement and Dowload Windows x64 compatible exe
4. Click the exe and change the
   1. C:\Program Files target folder to C:\jdk
   2. This is needed because there is some issue with dealing with space in “Program Files”

**Installing Apache Spark**

1. Go to <https://spark.apache.org/>
2. Click “Download Spark” button
3. Download “[spark-2.4.3-bin-hadoop2.7.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-2.4.3/spark-2.4.3-bin-hadoop2.7.tgz)” as of Jul 2nd, 2019
4. For extracting files for .tgz, we need WINRAR on Windows
   1. Download “Winrar x64” exe from <https://rarlab.com/download.htm>
   2. Click it and install it
5. Extract the files using WinRAR from .tgz
6. Create C:\spark and copy all the extracted files to that directory
7. Go to C:\spark\conf
   1. Change the filename “log4j.properties.template” to “log4j.properties”
   2. Open the file and change “log4j.rootCategory=ERROR, console” to “log4j.rootCategory=INFO, console”
   3. Save the file

**Don’t install HADOOP but make sure Spark thinks HADOOP is there by doing the below on WINDOWS**

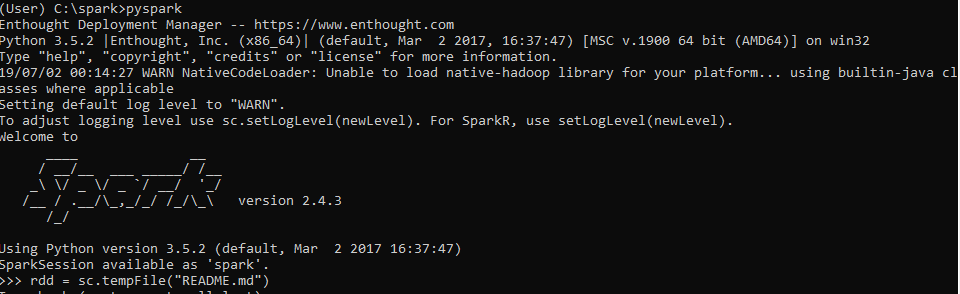
1. Download winutils.exe
2. Create a directory C:\winutils\bin
3. Copy winutils.exe to C:\winutils\bin
4. Create C:\tmp\hive directory
5. Go to Command Prompt and do
   1. cd C:\winutils\bin
   2. winutils.exe chmod 777 C:\tmp\hive

**Set Environment Variables**

1. Go to Control Panel -> System and Security -> System -> Advanced System Settings -> Environment Variables
2. Set the ENVs as below:
   1. SPARK\_HOME = C:\spark
   2. JAVA\_HOME = C:\jdk
   3. HADOOP\_HOME = C:\winutils
   4. Add the below in PATH
      1. %SPARK\_HOME%\bin
      2. %JDK\_HOME%\bin

**Running pyspark**

1. Double Click “Enthought Canopy”
2. Click Tools -> Canopy Command Prompt
3. cd C:\spark
4. pyspark



Executing a sample program which counts number of lines in a file



‘sc’ here in the above program is “SparkContext object”

‘pyspark’ is the Python API to avail Spark Services in Python

‘rdd’ means ‘distributed dataset’

**Get some Movie Ratings Data Set**

1. Go to ‘grouplens.org”
2. Download “MovieLens 100k Dataset” from there

**How to execute a pyspark script?**

1. Use ‘spark-submit’ as below to execute a pyspark script:

