

**Assignment 1**

**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**

Python, Hadoop and Pyspark installation manual

Contents

[1. Installing Python 3](#_Toc528385874)

[2. Installation setup 7](#_Toc528385875)

[Java JDK installation guide 7](#_Toc528385876)

[Hadoop 2.8.0 installation guide 9](#_Toc528385877)

[Environment setting (system variable) 11](#_Toc528385878)

[Hadoop configuration 13](#_Toc528385879)

[Test the installation 18](#_Toc528385880)

[3. Install pyspark 22](#_Toc528385881)

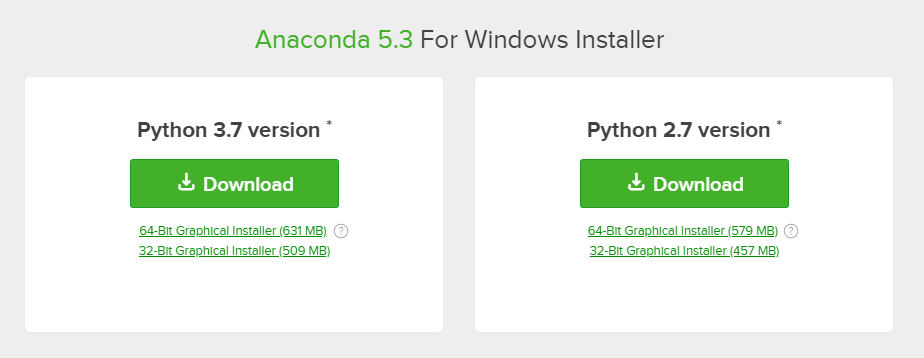
# Installing Python

Python can be installed easily by using ANACONDA. ANACONDA is the most popular Python/R data science distribution where you can manage your packages, dependencies and environments easily. It is free and open source.

1. To get started with ANACONDA, download it from below link

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

This instruction will use Windows as the OS platform so we choose the installer for Windows.



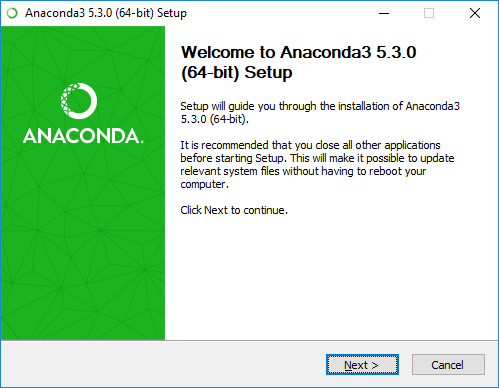
\*Note that you need to check for compatibility for the version of python that you want to install.

In this manual, we will download the latest Python version which is 3.7. Click on the download button.

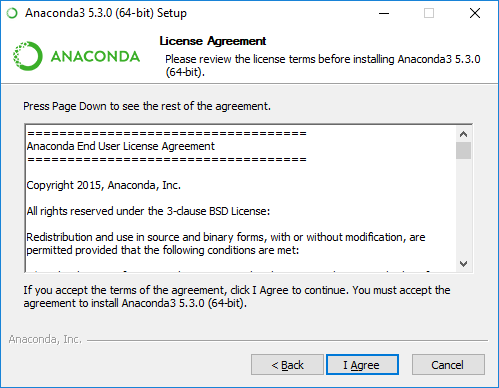
1. Once the installer has been downloaded, double click on the installer to start the installation.



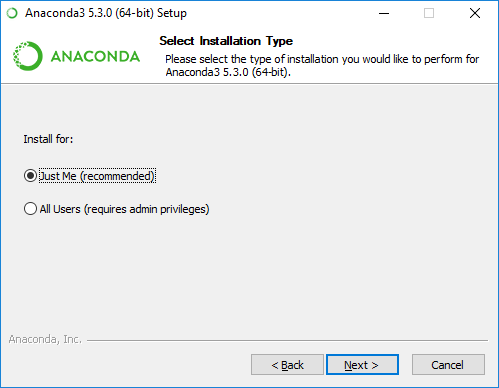
1. Click on “Next”.



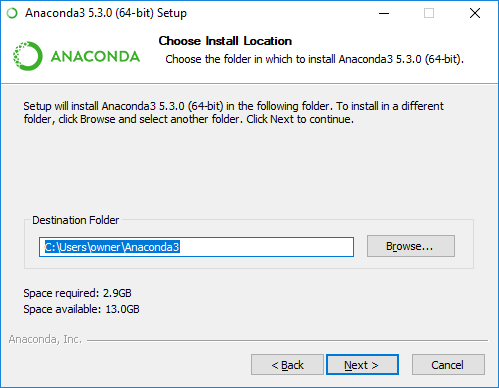
1. Click on “Agree” button.



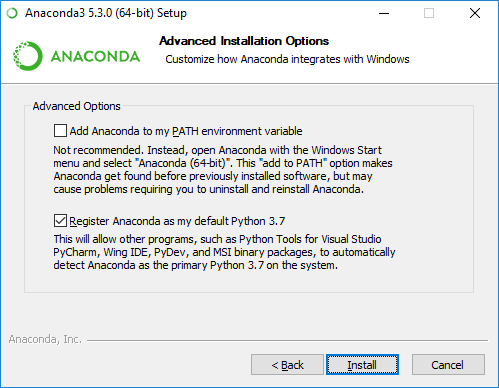
1. Select the installation type that you want and click “Next”



1. Select the installation location and click “Next”



1. Select “Register Anaconda as my default Python 3.7” and click “Install” button



Since Python 3.7 is relatively new, there might be some compatibility issues. In that case you maybe need to use Python 3.5 or 3.6. Follow the instruction in below link to get Anaconda with Python 3.5 or 3.6.

<http://docs.anaconda.com/anaconda/faq/#how-do-i-get-anaconda-with-python-3-5-or-3-6>

# Installation setup

To Install Hadoop on Window 10 64-bit, the first steps are to download and install these software.

1. Download and install Hadoop 2.8.0 from below link

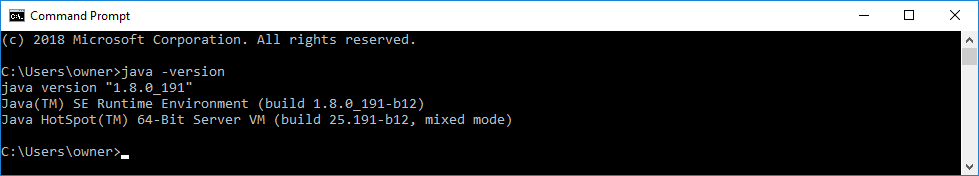
<http://archive.apache.org/dist/hadoop/core//hadoop-2.8.0/hadoop-2.8.0.tar.gz>

1. Download and install Java JDK from below link

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

## Java JDK installation guide

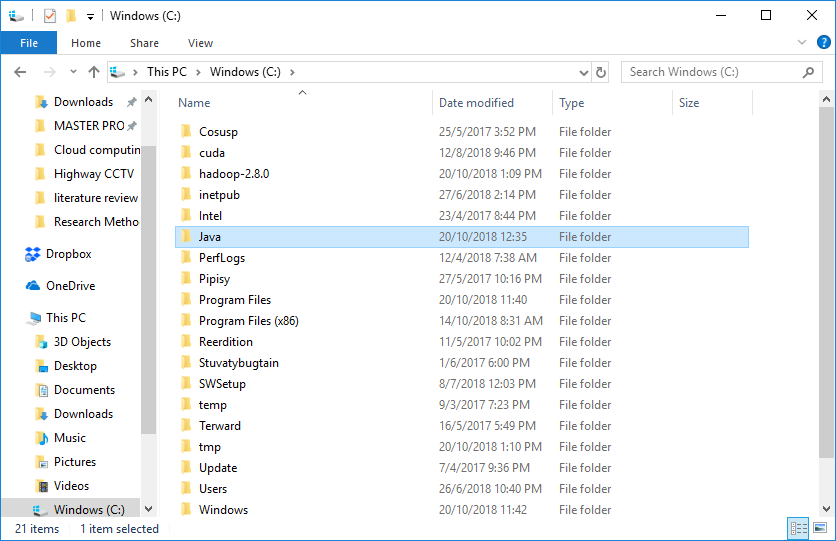
1. Check if Java JDK has been installed on your computer or not. Open CMD prompt and type “java -version” to check for the status.



1. If Java is already installed, check where is the installed is located. If it is located inside Program Files folder, copy or move the installed Java folder to “C:\” folder[[1]](#footnote-1).

The new Java path should be as below.

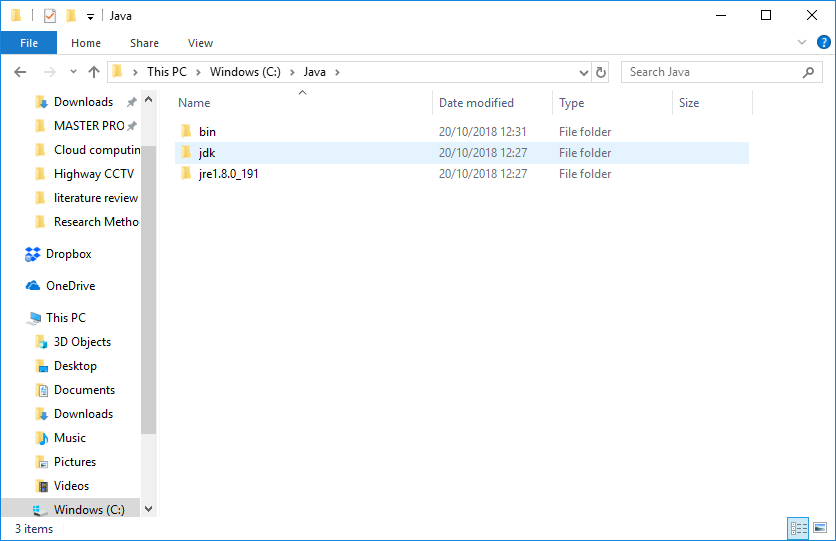
“C:\Java”



You can also rename the “jdk1.8.0\_version number” folder to “jdk”

The new java jdk folder path should be as below

“C:\Java\jdk”



1. If java is not installed on your system, then install the java to “C:\Java” as in step 2).

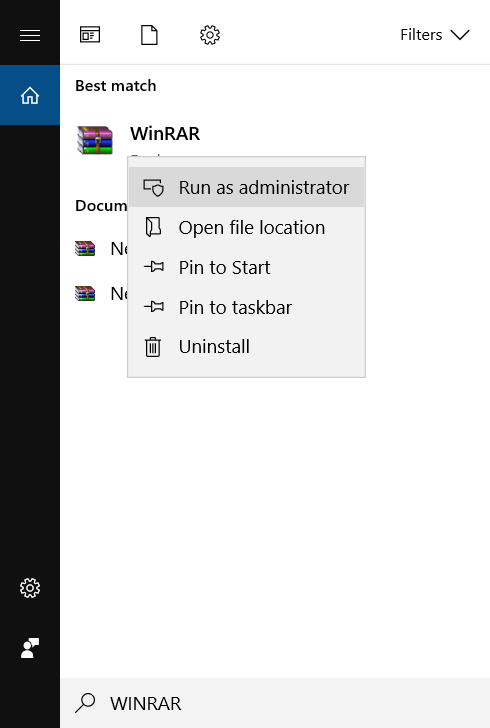
## Hadoop 2.8.0 installation guide

1. Extract file Hadoop 2.8.0.tar.gz or Hadoop-2.8.0.zip and place under "C:\Hadoop-2.8.0"

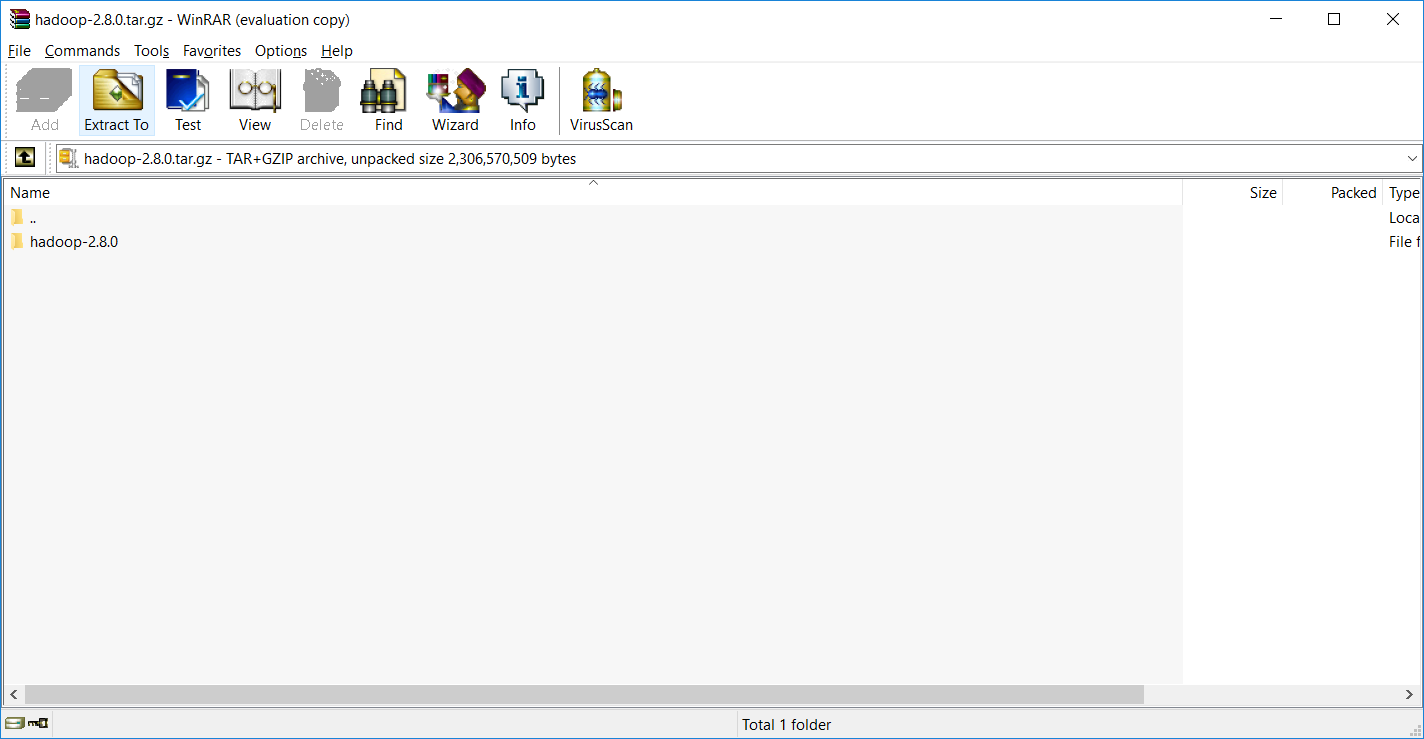
To extract the file, you need to run the extraction software with Administrator privilege.

Below example is how you extract the file using WINRAR.

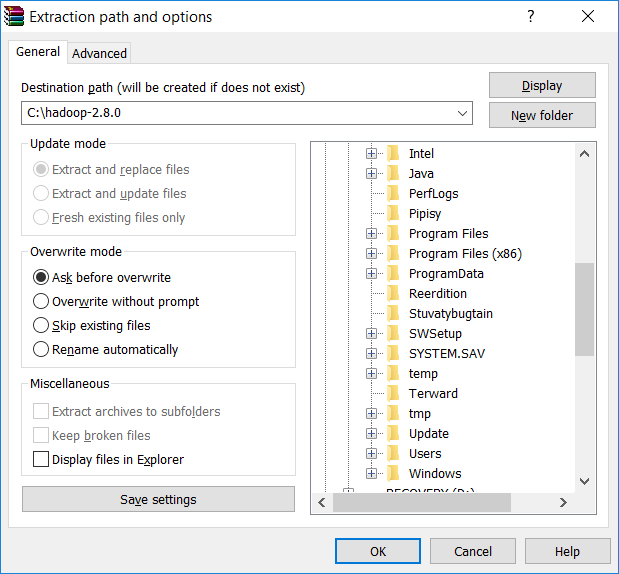
1. Open and Run as Administrator

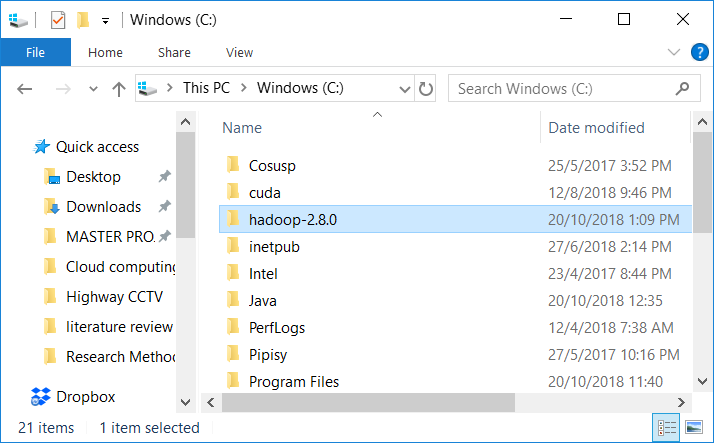


1. Select Hadoop installation file



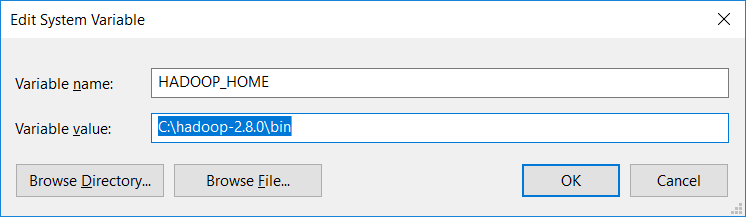
1. Select installation path (install in C drive)



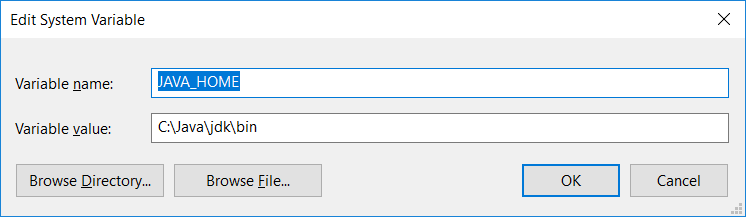


## Environment setting (system variable)

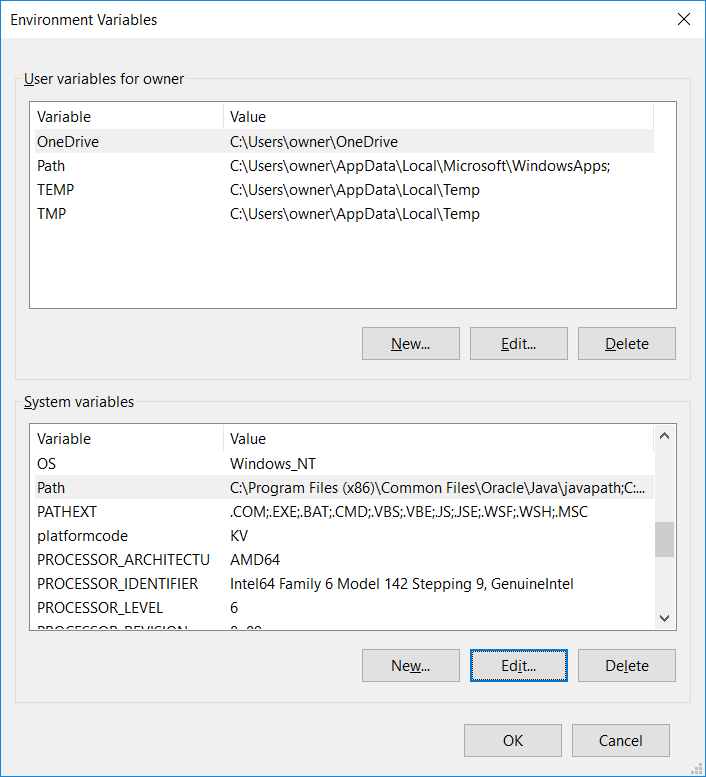
1. Set the path HADOOP\_HOME Environment variable on windows 10 (create new variable)

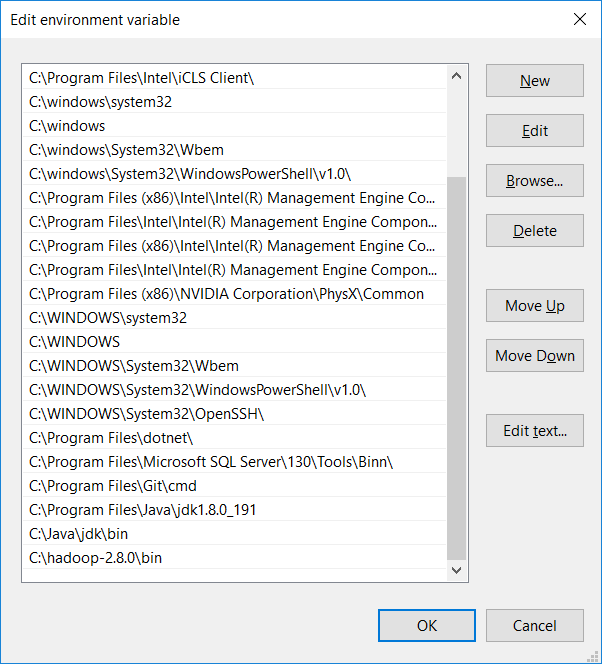


1. Set the path JAVA\_HOME Environment variable on windows 10 (create new variable)



1. Add binary file path from Hadoop and Java to system path variable (edit path variable)





## Hadoop configuration

1. Edit file C:/Hadoop-2.8.0/etc/hadoop/core-site.xml, paste below xml paragraph and save this file.

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

1. Rename "mapred-site.xml.template" to "mapred-site.xml" and edit this file C:/Hadoop-2.8.0/etc/hadoop/mapred-site.xml, paste below xml paragraph and save this file

<configuration>

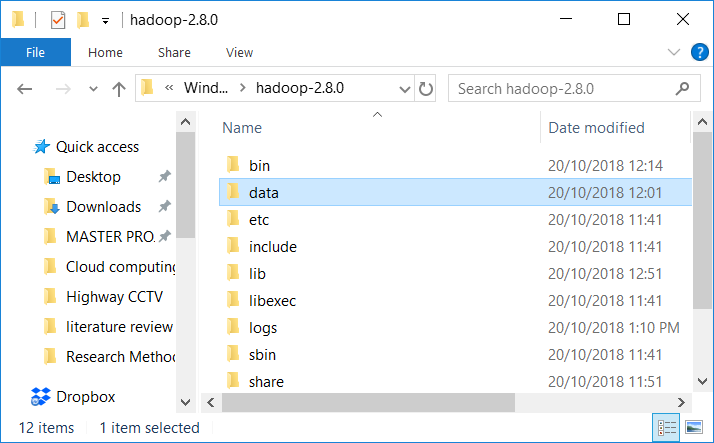
<property>

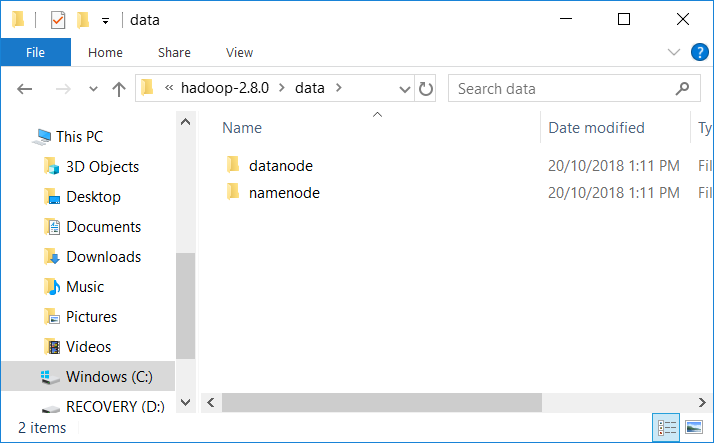
<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

1. Create folder "data" under "C:\Hadoop-2.8.0"
2. Create folder "datanode" under "C:\Hadoop-2.8.0\data"
3. Create folder "namenode" under "C:\Hadoop-2.8.0\data"



1. Edit C:\hadoop-2.8.0\etc\hadoop\hdfs-site.xml, paste below xml paragraph and save this file.

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>C:\hadoop-2.8.0\data\namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>C:\hadoop-2.8.0\data\datanode</value>

</property>

</configuration>

1. Edit file C:/Hadoop-2.8.0/etc/hadoop/yarn-site.xml, paste below xml paragraph and save this file.

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

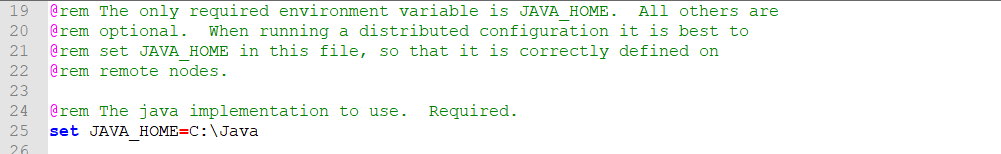
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

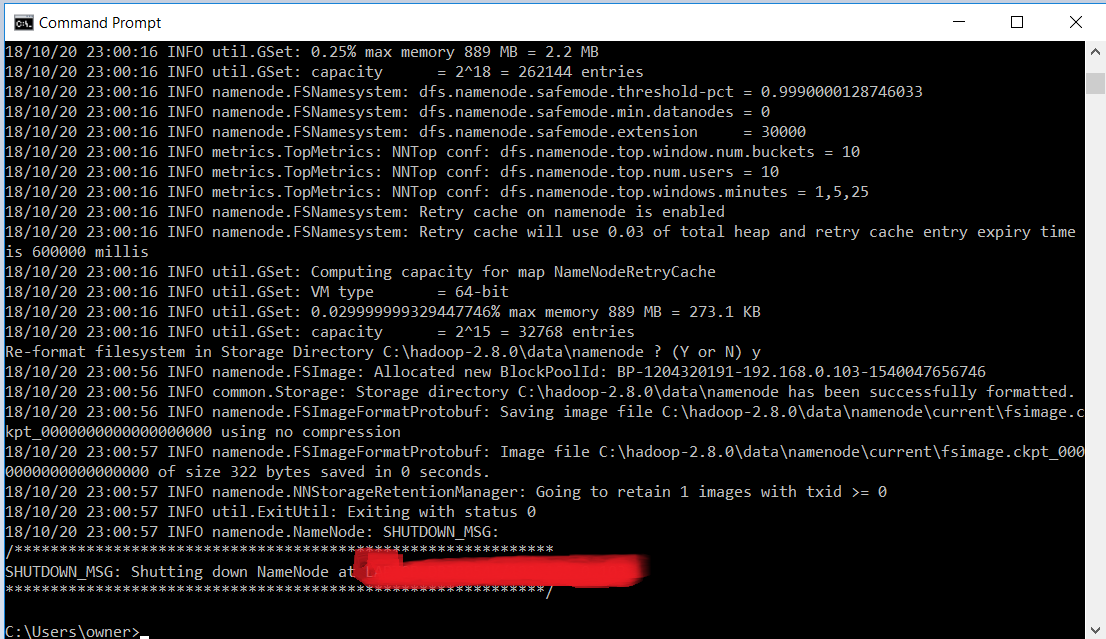
1. Edit file C:/Hadoop-2.8.0/etc/hadoop/hadoop-env.cmd by replacing the command line "JAVA\_HOME=%JAVA\_HOME%" with "JAVA\_HOME=C:\Java" instead(On C:\java this is path to file jdk.18.0)



1. Dowload file Hadoop Configuration.zip

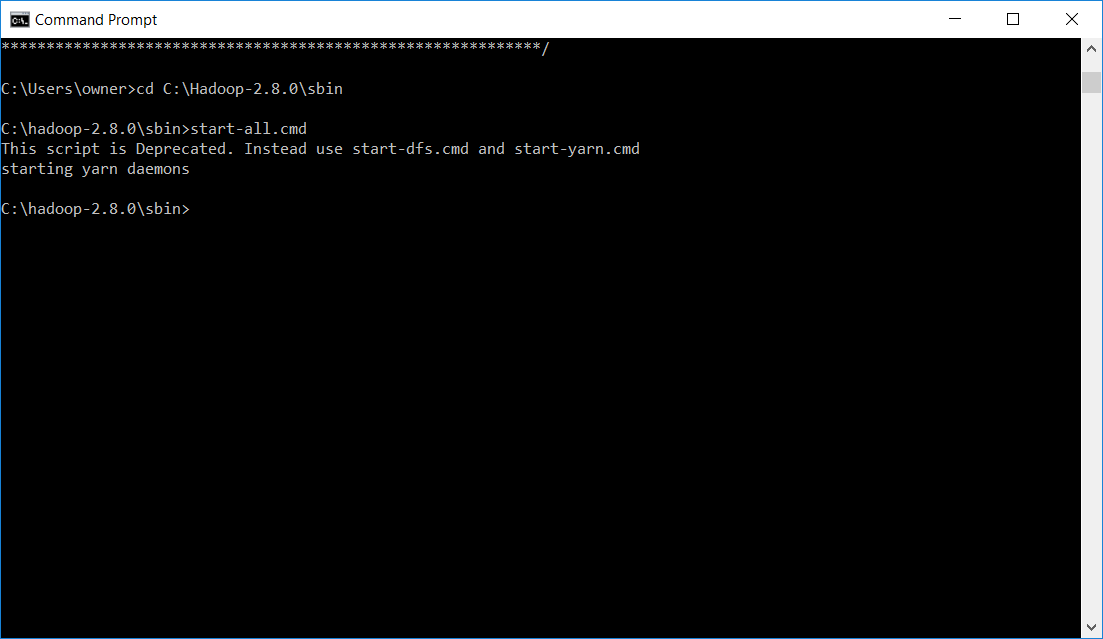
<https://github.com/Nidzam81/hadoop-setup/blob/master/Hadoop%20Configuration.zip>

1. Delete file bin on C:\Hadoop-2.8.0\bin, replaced by file bin on file just download (from Hadoop Configuration.zip)
2. Open cmd and type command "hdfs namenode –format" .

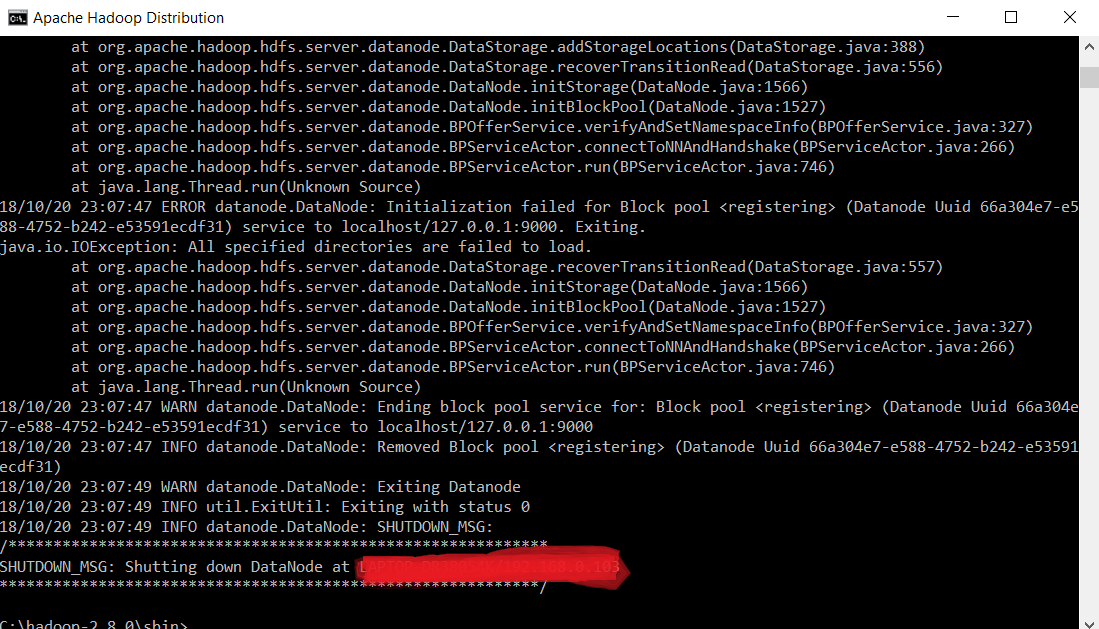


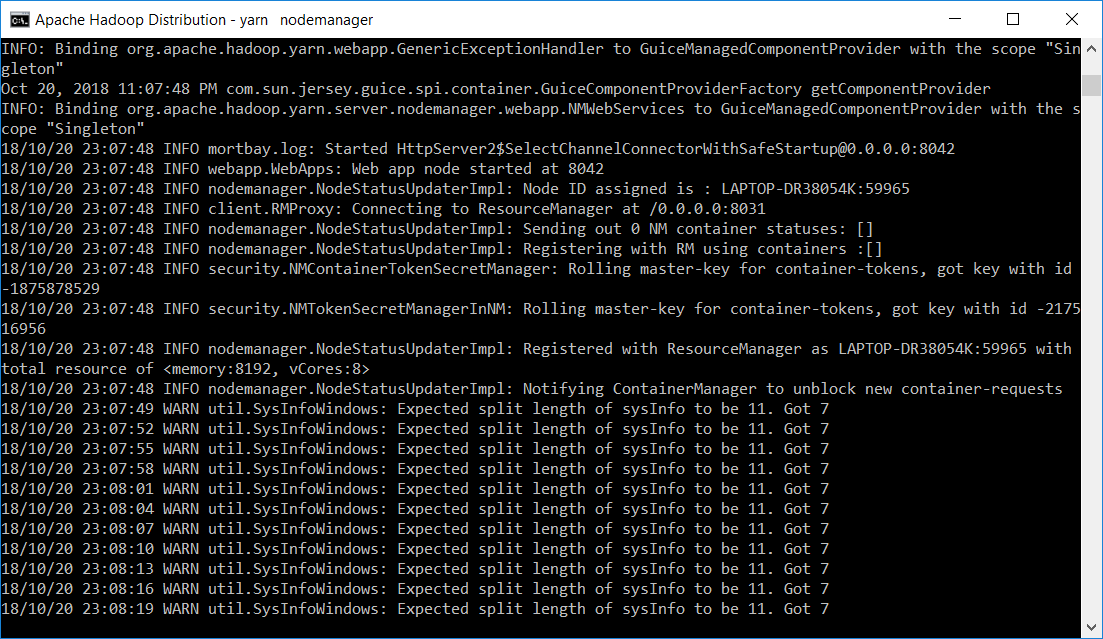
## Test the installation

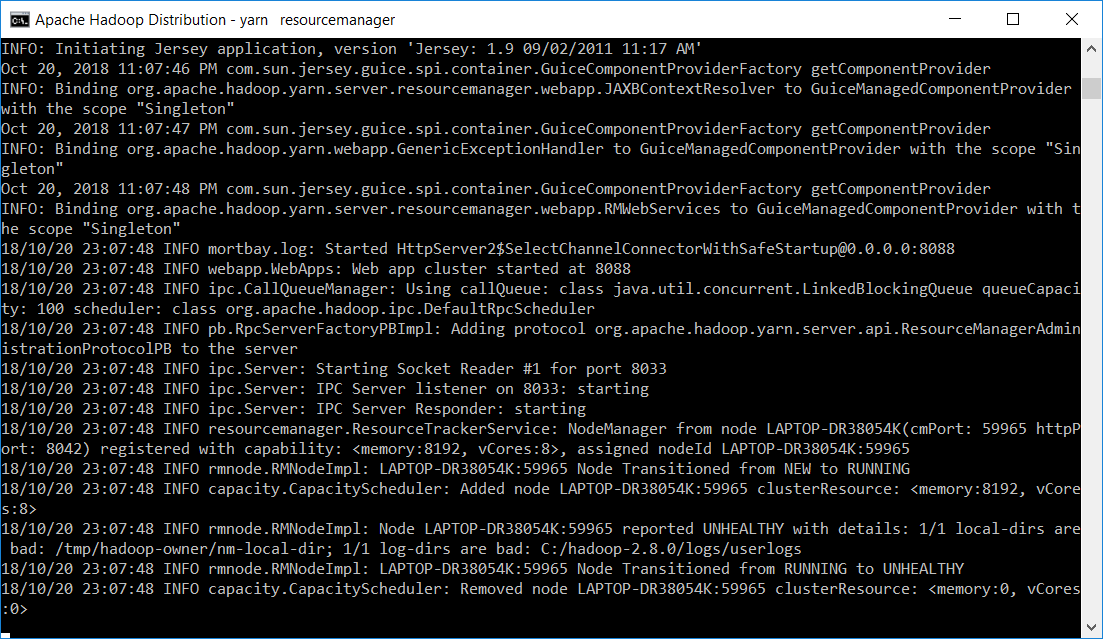
1. Open cmd and change directory to "C:\Hadoop-2.8.0\sbin" and type "start-all.cmd" to start apache.

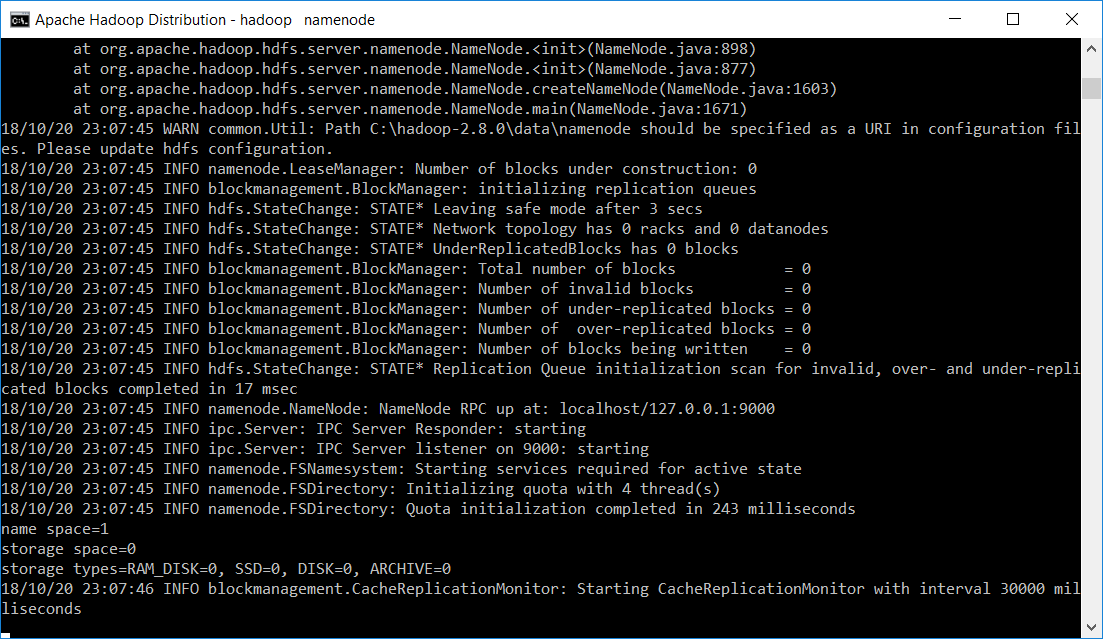


1. Make sure these apps are running
   1. Hadoop Namenode
   2. Hadoop datanode
   3. YARN Resourc Manager
   4. YARN Node Manager

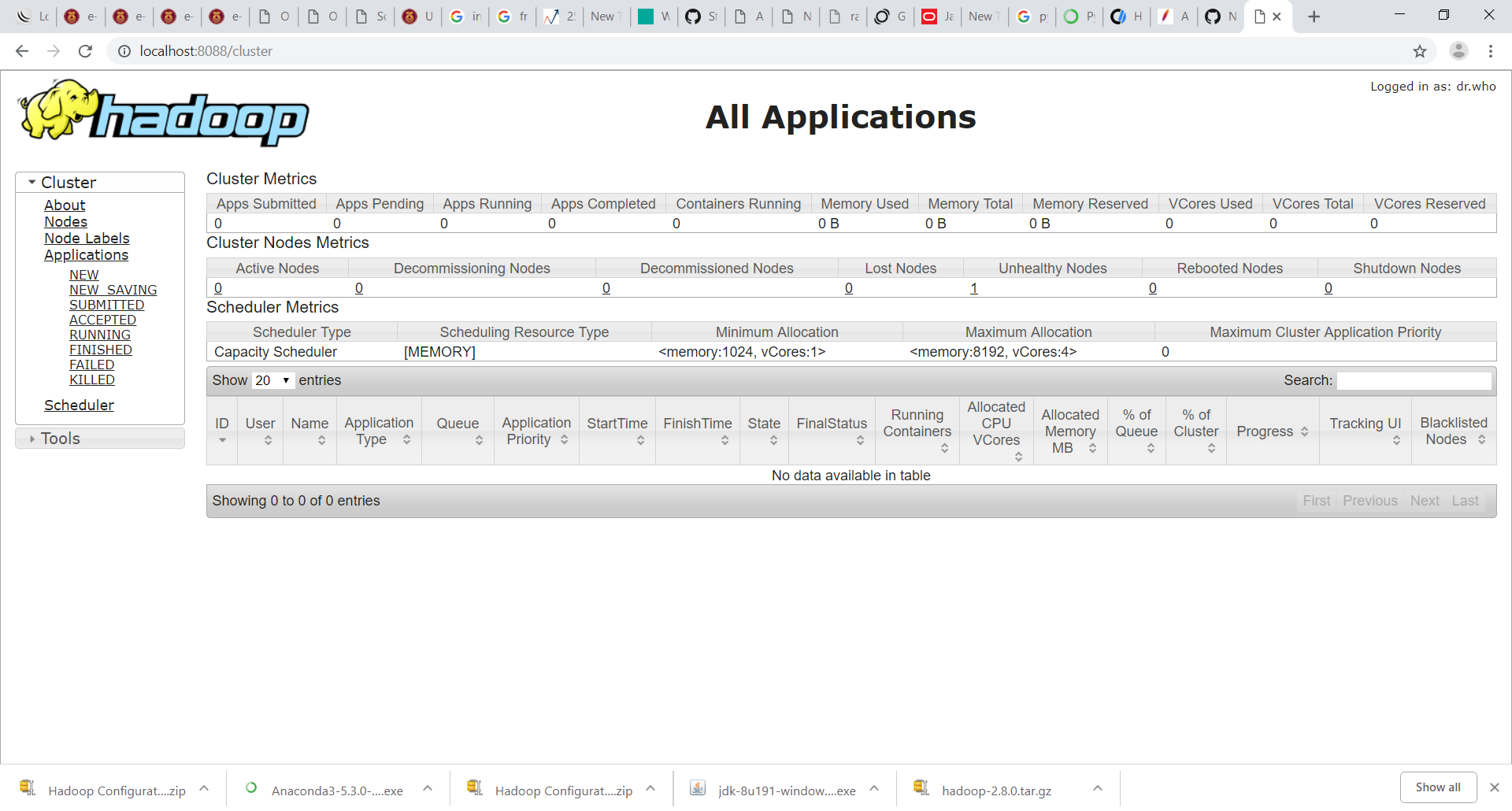




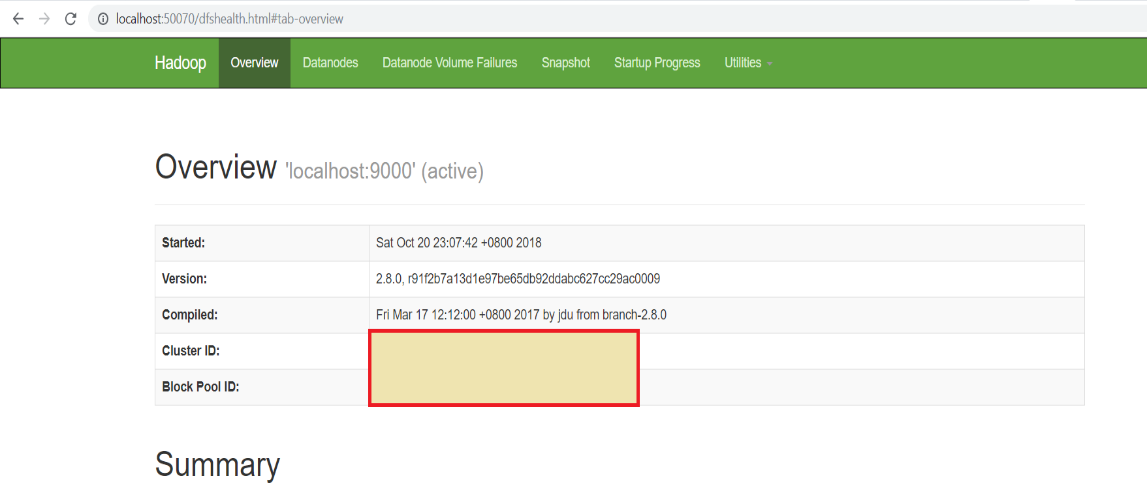




1. Open: [http://localhost:8088](http://localhost:8088/)



1. Open: <http://localhost:50070>

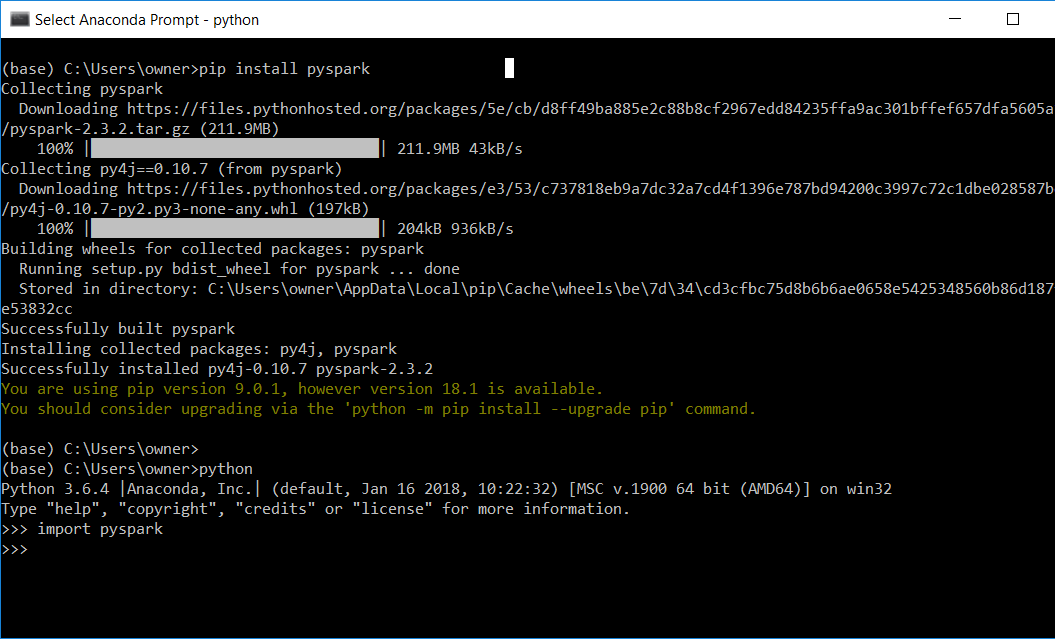


# Install pyspark

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



1. The reason we need to do this is because the installation will not recognize the “Program Files” folder path due to the spaces between the folder name. [↑](#footnote-ref-1)