Build Machine learning Lab



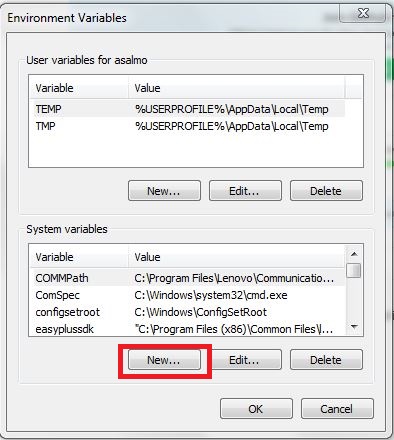
* Java (OpenJDK 8)
* Hadoop (using version: 3.1.2)
* Spark (using version: 2.7.0 to 3.2.0 preview)
* Anaconda (Jupyter). Windows 7: 2018 and Windows 10: 2019

1. If you don’t have Java 8, you will need to implement step 1 & 2. If you already have Java 8, you can pass step # 1. Hadoop 3.1.2 works with Java 8.

Install OpenJDK-8 Java

Download and install (<https://developers.redhat.com/products/openjdk/download?extIdCarryOver=true&sc_cid=701f2000001OH7JAAW>)

1. Add JAVA\_HOME to environment variables.



JAVA\_HOME= C:\alaa\AdoptOpenJDK\jdk-8.0.232.09-hotspot

Add JAVA\_HOME to path %JAVA\_HOME%\bin

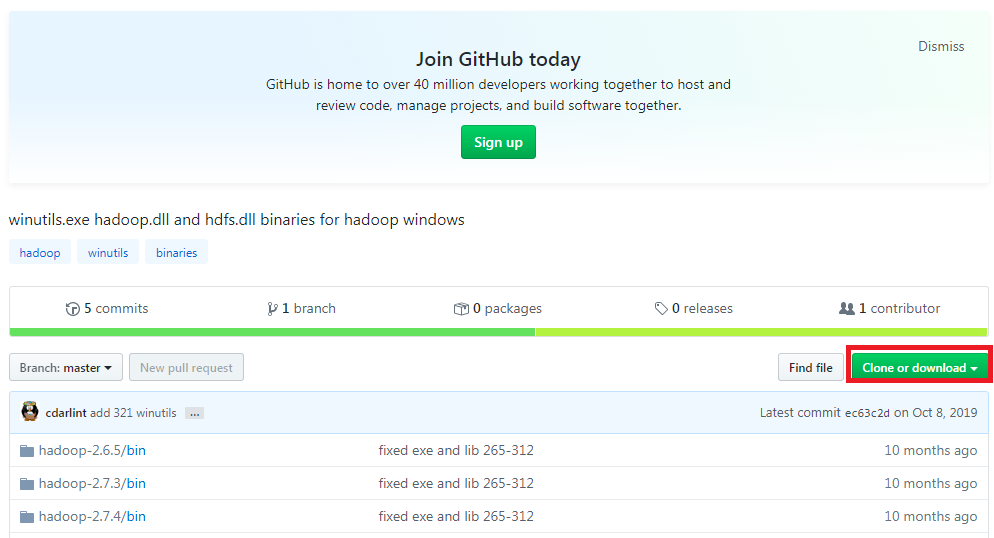
1. Build Hadoop on Windows (one node)
2. Download and install <https://www.7-zip.org/> (to unzip Linux)
3. Download <http://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz>
4. Make directory “bigdata” on c drive

C:\bigdata

1. Unzip hadoop-3.1.2.tar.gz

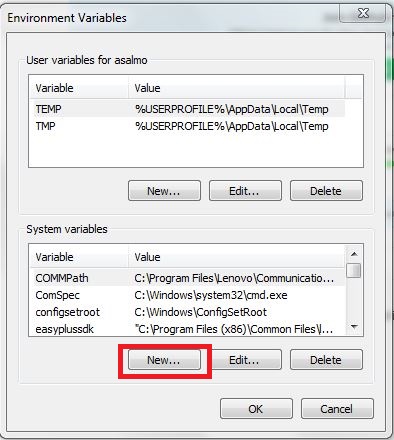
C:\bigdata\hadoop-3.1.2

1. Download Hadoop windows patch <https://github.com/cdarlint/winutils>



1. Download all patches, unzip the folder and copy Hadoop-3.1.2/bin to bin folder (C:\bigdata\hadoop-3.1.2\bin)
2. Build Hadoop variables:

Go “System properties” 🡪 Choose “Environment variables”



1. Press “New” to add:

* HADOOP\_HOME = C:\bigdata\ hadoop-3.1.2
* HADOOP\_BIN = C:\bigdata\ hadoop-3.1.2\bin
* HADOOP\_SBIN= C:\bigdata\ hadoop-3.1.2\sbin
* Add % HADOOP\_HOME %; % HADOOP\_BIN %;% HADOOP\_SBIN% to **path** variable

1. Configure Hadoop to run on single machine

We will need to change the following files (C:\ bigdata\hadoop-2.9.1\etc\hadoop):

hadoop-env.cmd

core-site.xml

hdfs-site.xml

mapred-site.xml

1. hadoop-env.cmd:

Change JAVA\_HOME variable

From:

set JAVA\_HOME=%JAVA\_HOME%

To:

set JAVA\_HOME=C:\AdoptOpenJDK\jdk-8.0.232.09-hotspot

1. core-site.xml

Open this file to add:

<configuration>

<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>

Note: Replace the empty < configuration></ configuration>

1. hdfs-site.xml

Open this file to add:

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value>file:///C:/bigdata/hadoop-3.1.2/data/namenode</value>

</property>

<property>

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

<value>file:///C:/bigdata/hadoop-3.1.2/data/datanode</value>

</property>

</configuration>

Note: Replace the empty < configuration></ configuration>

In this case, we will need to make directory for

C:\BigData\ hadoop-3.1.2\**data**

C:\BigData\ hadoop-3.1.2\data\**namenode**

C:\BigData\ hadoop-3.1.2\data\**datanode**

1. mapred-site.xml

Open this file to add:

<configuration>

<property>

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

<value>yarn</value>

</property>

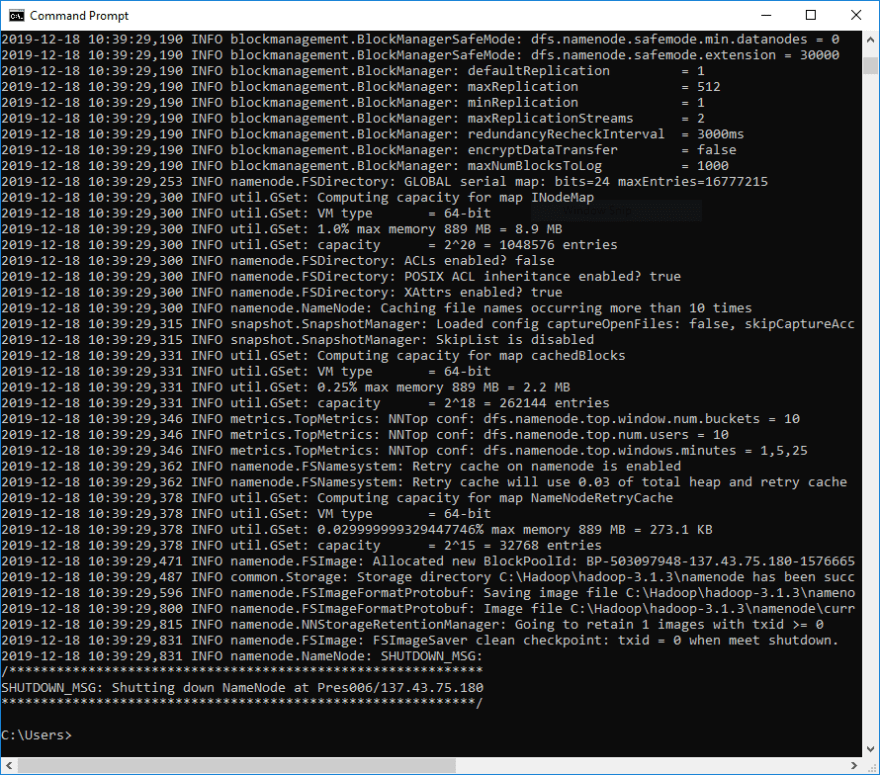
</configuration>

Note:

1. Replace the empty < configuration></ configuration>
2. Formant NameNode:

-Open cmd

-Type “hadoop namenode -format”



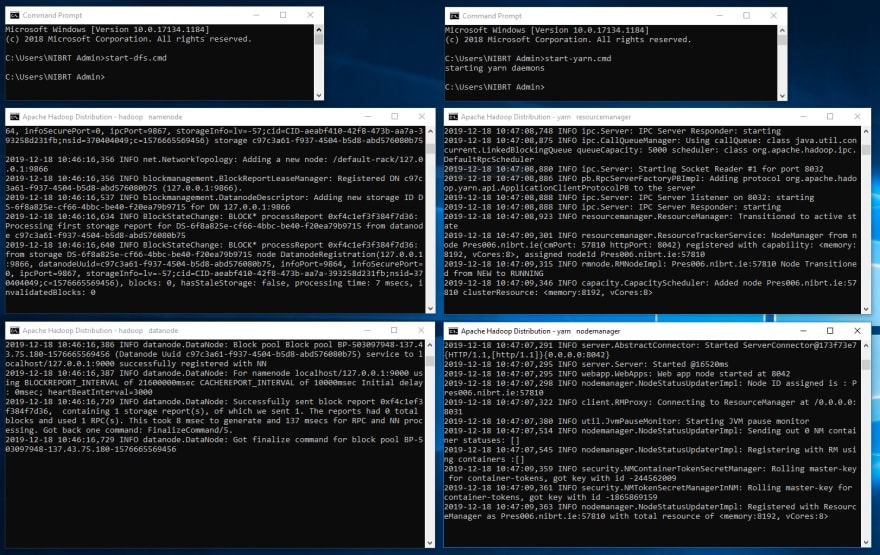
1. One more thing to do: copy hadoop-yarn-server-timelineservice-3.1.2

**from** C:\bigdata\hadoop-3.1.2\share\hadoop\yarn\timelineservice

**to** C:\bigdata\hadoop-3.1.2\share\hadoop\yarn

1. We need to type start-all.cmd to start all nodes on one machine

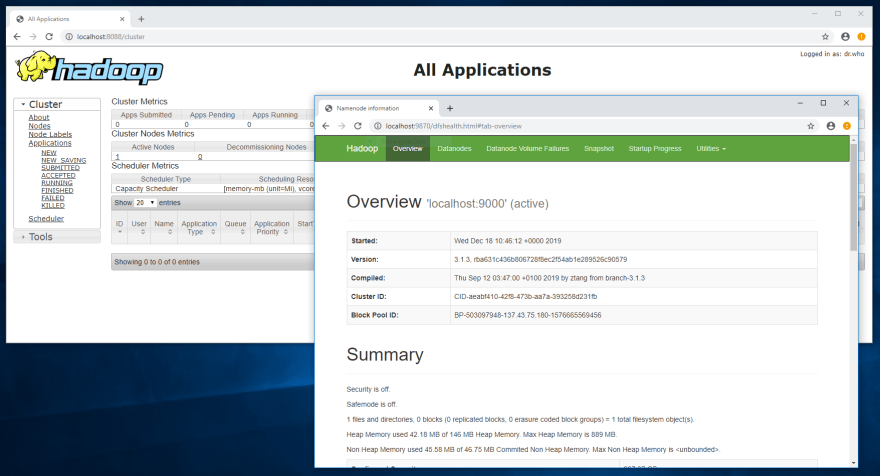
You should have:

[](https://res.cloudinary.com/practicaldev/image/fetch/s--p-uKiWJw--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/thepracticaldev.s3.amazonaws.com/i/hxs3insfi8dozupt41z3.PNG)

Now you should have Hadoop on your machine

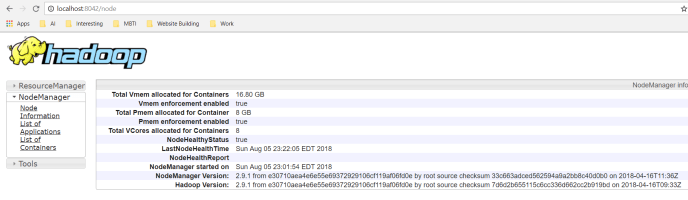
1. Hadoop Web UI

You can also open http://localhost:8088 and http://localhost:9870 in your browser



1. Node Manager

<http://localhost:8042>



K- Working with HDFS

>notepad Sample.txt

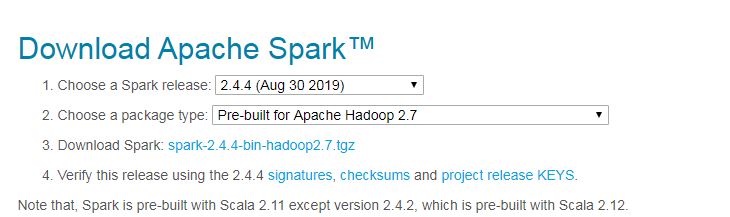
Write anything and save the file

hdfs dfs -ls /  
hdfs dfs -mkdir /test  
hdfs dfs -copyFromLocal Sample.txt /test  
hdfs dfs -ls /test  
hdfs dfs -cat /test/Sample.txt

1. Hadoop reference: <https://dev.to/awwsmm/installing-and-running-hadoop-and-spark-on-windows-33kc#comments>
2. Hadoop Patch reference <https://github.com/cdarlint/winutils>

Install Spark

1. Download Spark version 2.3 or 2.4 (<https://spark.apache.org/downloads.html>)



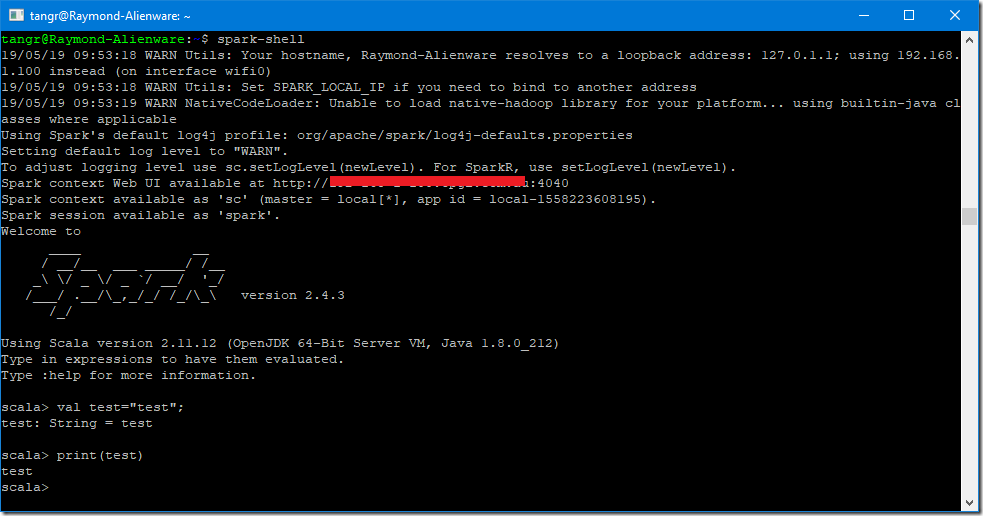
1. Unzip spark-2.4.4-bin-hadoop2.7.tgz
2. Put the folder in C:\bigdata\spark-2.4.4-bin-hadoop2.7
3. Add to system variables

SPARK\_HOME=C:\alaa\bigdata\spark-2.4.4-bin-hadoop2.7

PYSPARK\_PYTHON=C:\Users\asalmo\AppData\Local\Programs\Python\Python37\python.exe

PYSPARK\_DRIVER\_PYTHON= C:\Users\asalmo\AppData\Local\Programs\Python\Python37\python.exe

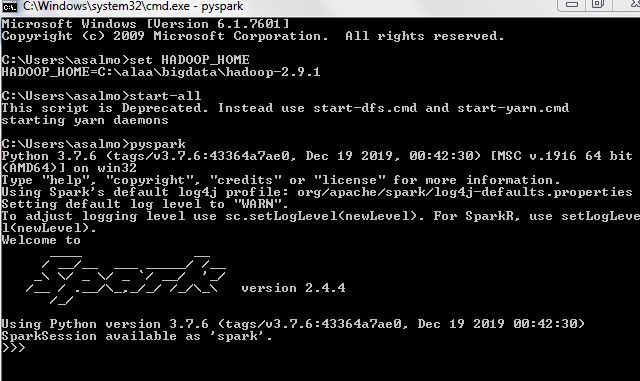
1. Add SPARK\_HOME to path % SPARK\_HOME %\bin
2. Open cmd
3. Type: spark-shell to start spark scala



1. Exit type :q
2. You need to install Python (download: <https://www.python.org/downloads/windows/>)
3. Add the variable: PYSPARK\_DRIVER\_PYTHON

PYSPARK\_DRIVER\_PYTHON= C:\Users\XXXXXXXX (Usename)\AppData\Local\Programs\Python\Python37\python.exe

1. Open cmd and type pyspark



1. To quit write quit()

Reference: <https://kontext.tech/column/spark/311/apache-spark-243-installation-on-windows-10-using-windows-subsystem-for-linux>

Anaconda with Jupyter notebook

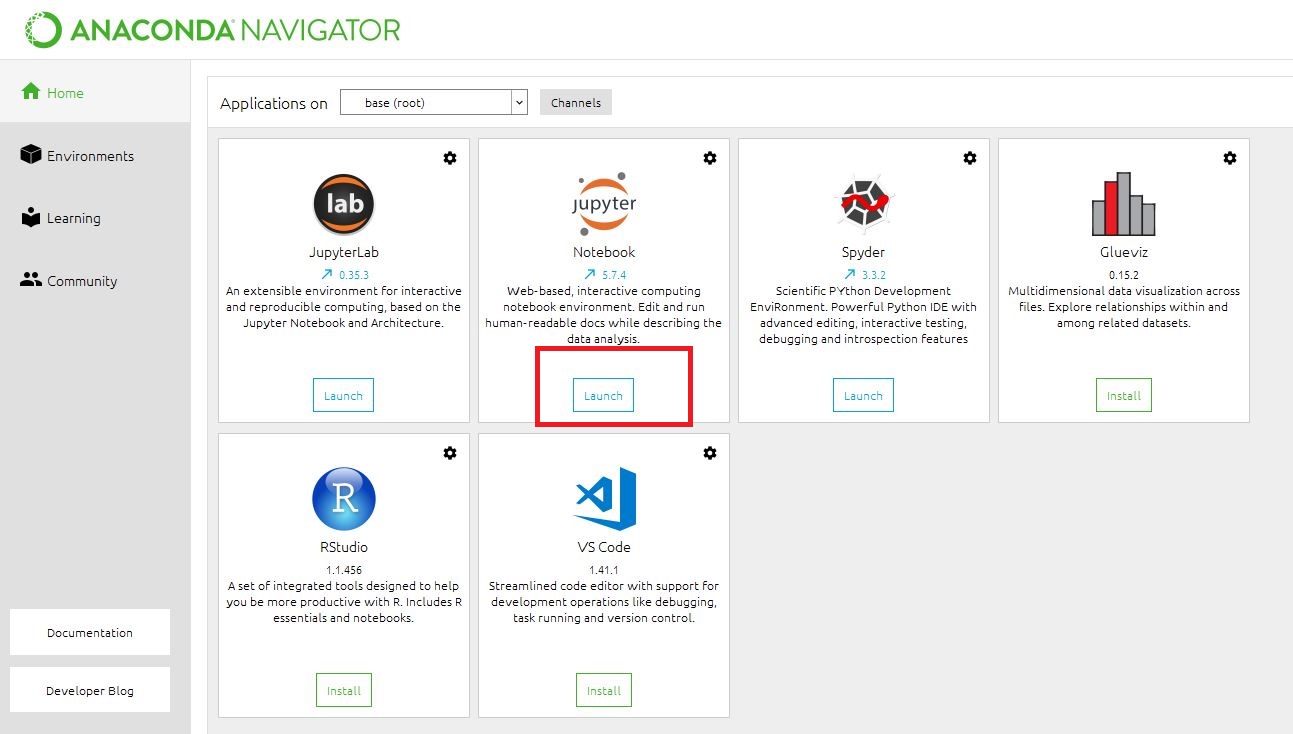
1. Install Anaconda

Note: For Windows 7 install Anaconda with 2018 and Python 3.7  
 For Windows 10 install Anaconda with 2019 and Python 3.7

<https://docs.anaconda.com/anaconda/packages/oldpkglists/>

1. After installation,

Go to window search for anaconda



Press Jupyter to start the web notebook.

1. Install findspark

Go to “search program and files” write “anaconda”

Choose “Anaconda Prompt”

Write "conda install -c conda-forge findspark"

1. Install Pandas

Go to “search program and files” write “anaconda”

Choose “Anaconda Prompt”

Wite "pip install pandas"