جامعة ابن طفيل
Université Ibn Tofaïl
Faculté des Sciences

Hadoop 3.2.2 Installation Guide (Windows) Documentation Revision v1.0

Karjout Abdeslam, Master Specialsé
Big Data & Cloud Computing

Contents

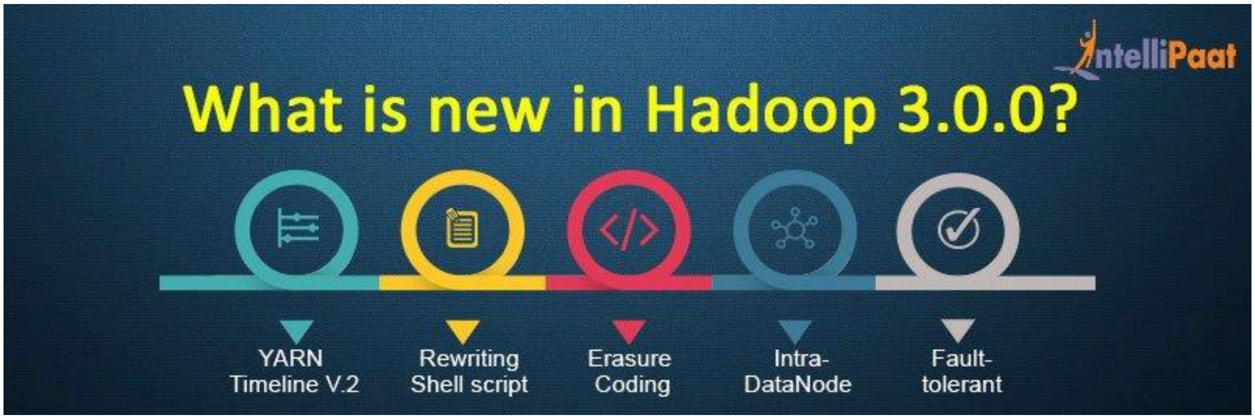
I. Introduction

This guide will assist you with installing Hadoop 3.2.2 into your own machine. This guide uses a 64---bit Windows 10 Pro . We recommend you use this version of Windows for best results.

As you will learn Hadoop is actually made up of several components:

- HDFS stand for (Hadoop Distributed File System)
- MapReduce
- YARN stand for (Yet Another Resource Negotiator)

It is easy to get these confused and some people will use them interchangeably. However, it is important to note that MapReduce is a programming paradigm and Hadoop provides framework that allows us to run MapReduce algorithms. Indeed, the inspiration for furthering MapReduce came out of a paper published by Google engineers. Thus began the Hadoop effort. Hadoop also includes the Hadoop Distributed File System, or HDFS. This also stems from a paper released from Google about the Google File System (GFS). These two components of Hadoop (MapReduce engine and HDFS) are probably the number one reason people use Hadoop in the first place. However, with the Hadoop 3.x release, a new component of Hadoop was announced: YARN (Yet Another Resource Negotiator). This allows for better scalability and cluster resource allocation – and also allows for applications to run on the Hadoop



infrastructure. These screenshot should help :

There are several ways to install Hadoop. One such way is through your favorite Linux distributions package---management system. However, you do not get much control over which versions of software you are installing. Simply installing Hadoop via a package manager masks much of Hadoop works internally. Another way is to use a distribution provided by Cloudera or Hortonworks. These tools provide cluster management services and a nice UI as well. In industry this would be the preferred option as you get a dashboard of your cluster status at all times. The last way is to install Hadoop manually. This allows you to see how the software works underneath (to an extent) and gives you a general idea as to how the other two installation methods actually work.

Note: Please be sure to follow these instructions EXACTLY. If you get an error on ANY of these steps, do not ignore it; if any of these goes wrong your Hadoop installation will not work. Hadoop is a distributed system with many moving parts. Even the smallest problem with your setup will cause Hadoop to break. Trust me, I've been there.

This guide was created with the following machine configuration*:*

---Windows 10 Pro

---8 CPUs, 16 GB RAM

---256 SSD & 1 To HDD

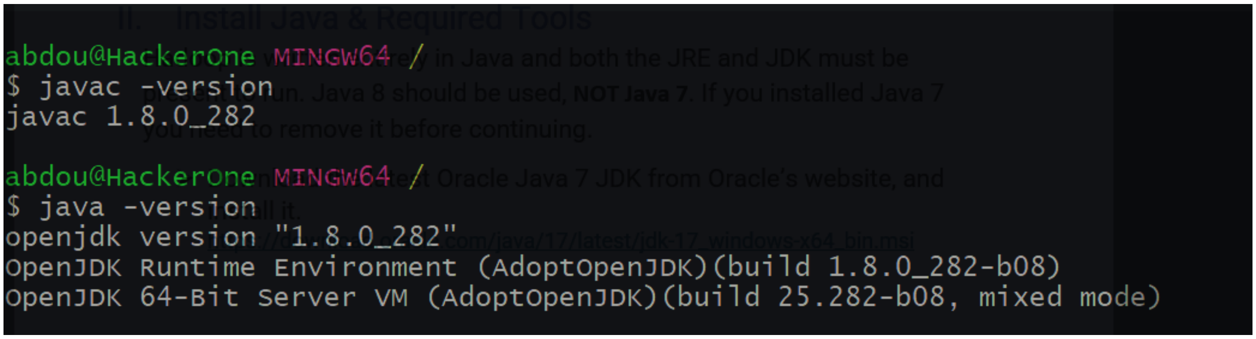
--- i7-7700HQ

Aside `– text in blue` is a file you must modify, `text in gray` is a command prompt screenshot.

II. Install Java & Required Tools

Hadoop is written entirely in Java and both the JRE and JDK must be present to run. Java 8 should be used, **NOT Java 7**. If you installed Java 7 you need to remove it before continuing.

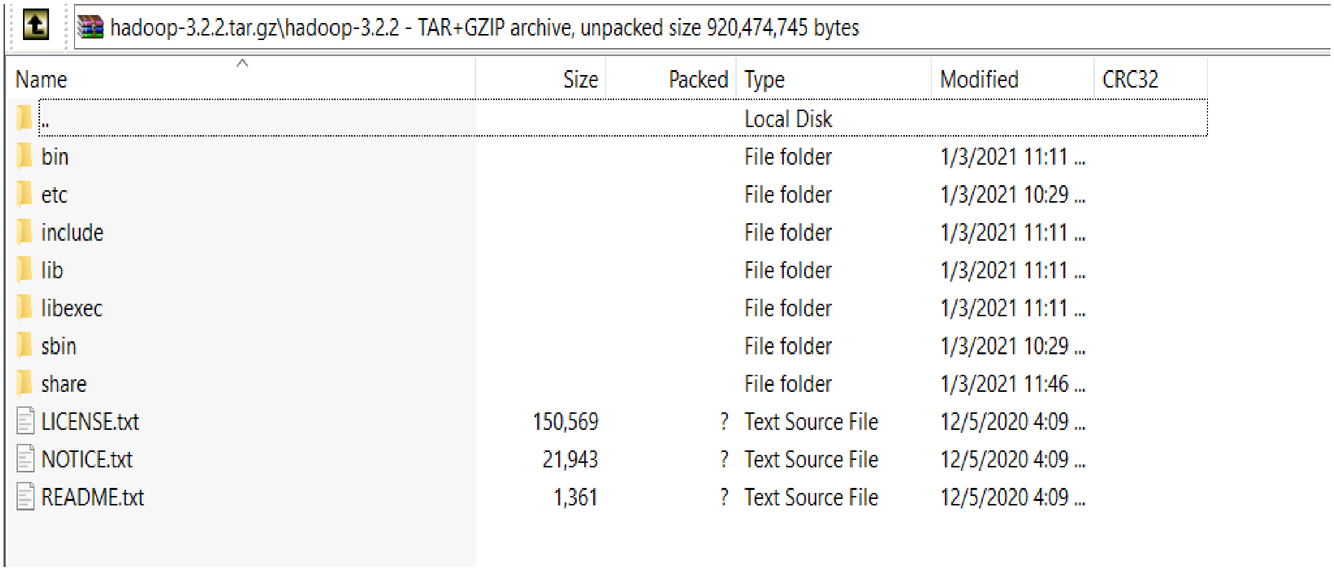
- ➢ Download the latest Oracle Java 7 JDK from Oracle's website, and install it.
  https://download.oracle.com/java/17/latest/jdk-17_windows-x64_bin.msi
  after downloading java 8 make sur if successfully installed



```
abdou@HackerOne MINGW64 /
$ javac -version
javac 1.8.0_282

abdou@HackerOne MINGW64 /
$ java -version
openjdk version "1.8.0_282"
OpenJDK Runtime Environment (AdoptOpenJDK)(build 1.8.0_282-b08)
OpenJDK 64-Bit Server VM (AdoptOpenJDK)(build 25.282-b08, mixed mode)
```

- ➢ Second You need to install an IDE in my case I will use Vscode
  link : https://code.visualstudio.com/download
- ➢ Download the Hadoop Source Tar V.3.2.2, We'll extract it later.
  Link : https://hadoop.apache.org/releases.html



hadoop-3.2.2.tar.gz\hadoop-3.2.2 - TAR+GZIP archive, unpacked size 920,474,745 bytes

Name	Size	Packed	Type	Modified	CRC32
..			Local Disk		
bin			File folder	1/3/2021 11:11 ...	
etc			File folder	1/3/2021 10:29 ...	
include			File folder	1/3/2021 11:11 ...	
lib			File folder	1/3/2021 11:11 ...	
libexec			File folder	1/3/2021 11:11 ...	
sbin			File folder	1/3/2021 10:29 ...	
share			File folder	1/3/2021 11:46 ...	
LICENSE.txt	150,569	?	Text Source File	12/5/2020 4:09 ...	
NOTICE.txt	21,943	?	Text Source File	12/5/2020 4:09 ...	
README.txt	1,361	?	Text Source File	12/5/2020 4:09 ...	

- Due to User Account Control, you will need to take ownership of the directories you create on the C:\ drive. To do this, open a Command Prompt as an Administrator. To do this, type in cmd in the Windows Search bar. Then right click on Command Prompt and choose “Run as Administrator.” Then type in the commands below. We’ll be using the directories later.
- Now Create a new folder in my case “hadoop_bigdata” using these command

```
C:\>mkdir hadoop_bigdata
```

- Extract hadoop-3.2.2.tar.gz to Hadoop_bigdata folder then rename the Hadoop 3.2.2 to Hadoop

```
C:\>cd hadoop_bigdata

C:\hadoop_bigdata>ren hadoop-3.2.2 hadoop

C:\hadoop_bigdata>cd Hadoop

C:\hadoop_bigdata\hadoop>dir

 Directory of C:\hadoop_bigdata\hadoop

10/25/2021  04:20 PM    <DIR>          .

10/25/2021  04:20 PM    <DIR>          ..

10/25/2021  04:20 PM    <DIR>          bin

10/25/2021  04:20 PM    <DIR>          etc

10/25/2021  04:20 PM    <DIR>          include

10/25/2021  04:20 PM    <DIR>          lib

10/25/2021  04:20 PM    <DIR>          libexec

12/05/2020  04:09 PM           150,569 LICENSE.txt

12/05/2020  04:09 PM            21,943 NOTICE.txt

12/05/2020  04:09 PM             1,361 README.txt

10/25/2021  04:20 PM    <DIR>          sbin
```

- ➢ Setting 777 permissions to a file or directory means that it will be readable, writable and executable

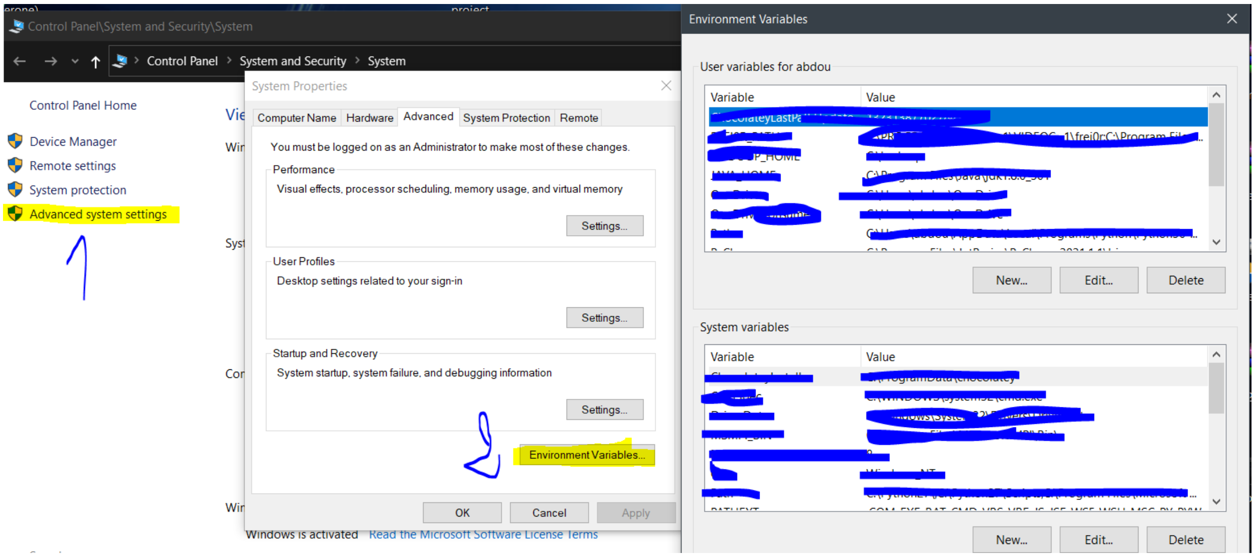
```
C:\hadoop_bigdata\hadoop>winutils.exe chmod 777 C:\hadoop_bigdata\hadoop
```

Now the Hadoop folder own all the permissions

III. Set Environment Variables :

Well, there’s two ways to set Environment Variables by using cmd or user interface so let’s choose the easiest one (GUI ☺XD!!!)

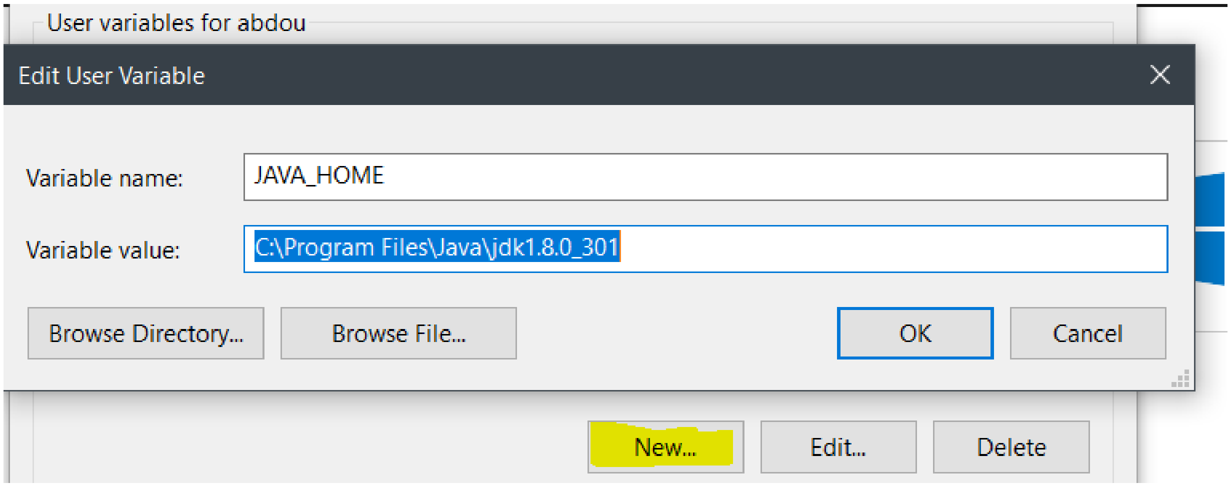
- ➢ Go to Control Panel>System>Advanced SystemSettings>Environment Variables:



- Add the following user environment variables. **Note that they are case---sensitive.**

  JAVA_HOME: C:\ " Your Jdk Path " in my case jdk path:

  C:\Program Files\Java\jdk1.8.0_301



User variables for abdou

Edit User Variable

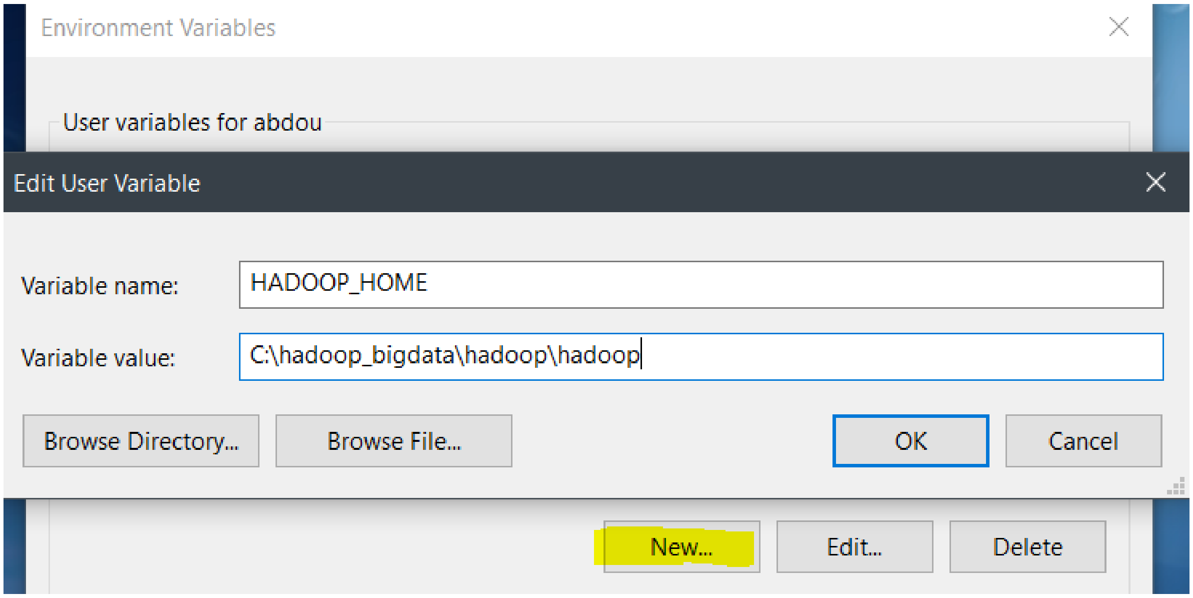
Variable name: JAVA_HOME

Variable value: C:\Program Files\Java\jdk1.8.0_301

Now the same thing for Hadoop

HADOOP_HOME : C:\"Your path to Hadoop " in my case

C:\hadoop_bigdata\hadoop



Environment Variables

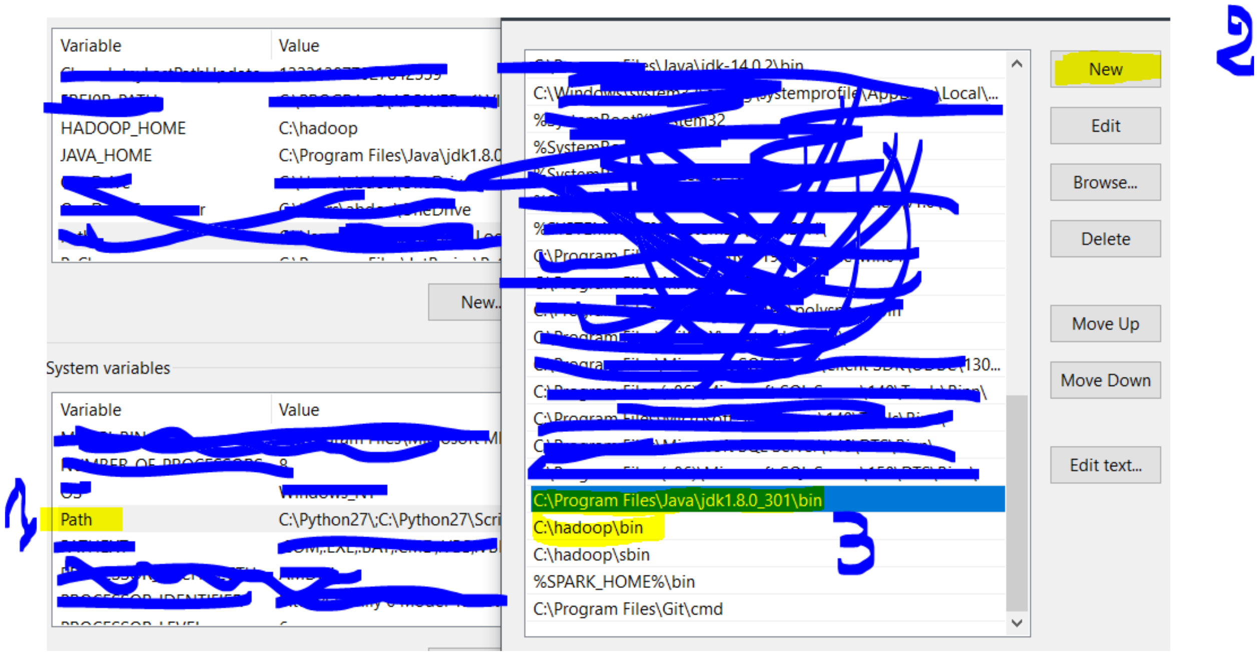
User variables for abdou

Edit User Variable

Variable name: HADOOP_HOME

Variable value: C:\hadoop_bigdata\hadoop\hadoop

- **CAREFULLY** edit the **System** environment variable **Path**, add the following to the END of the Variable value.
  add the following path to System environment variable path
  C:\Program Files\Java\jdk1.8.0_301\bin
  C:\hadoop_bigdata\hadoop\bin



Variable	Value
[redacted]	[redacted]
[redacted]	[redacted]
HADOOP_HOME	C:\hadoop
JAVA_HOME	C:\Program Files\Java\jdk1.8.0
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

System variables

Variable	Value
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
Path	C:\Python27\;C:\Python27\Scri
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]
C:\Program Files\Java\jdk1.8.0_301\bin
C:\hadoop\bin
C:\hadoop\sbin
%SPARK_HOME%\bin
C:\Program Files\Git\cmd

- Now let’s check if everything is OK

```
C:\Users\abdou>hadoop

Usage: hadoop [--config confdir] [--loglevel loglevel]
COMMAND

where COMMAND is one of:

  fs                   run a generic filesystem user client

  version              print the version

  jar <jar>            run a jar file

                       note: please use "yarn jar" to launch

                             YARN applications, not this command.

  checknative [-a|-h]  check native hadoop and compression
libraries availability

  conftest             validate configuration XML files

  distch path:owner:group:permisson
```

If you get this you are on the right way ☺

Wait !!! now the fun is begins

IV. Configure Hadoop

- First step locate to **C:\hadoop_bigdata\hadoop** using cmd and create folder called “**data**” then locate to C:\hadoop_bigdata\hadoop\data
- Create two folders under name **namenode** **&** **datanode**

```
C:\hadoop_bigdata\hadoop>mkdir data

C:\hadoop_bigdata\hadoop>cd data

C:\hadoop_bigdata\hadoop\data>mkdir namenode

C:\hadoop_bigdata\hadoop\data>mkdir datanode
```

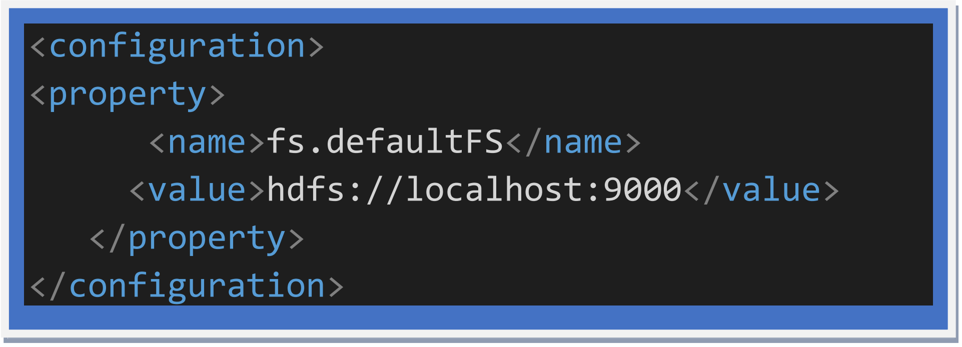
- Now that all installation directories are configured, we need to edit the Hadoop configuration files. *Use Vscode to edit the following files*. Add the following contents in between the <configuration> … </configuration> brackets in the files.

Note: Be careful when editing the contents of these files. You will encounter extra dashes/characters when copy—pasting from this file due to the PDF conversion. The easiest way is to either type it out carefully by hand or download the copies of the Config Files in the Google Drive (located here:

https://drive.google.com/drive/folders/1u-p_98_vORabi72LT4Ov9bumD8qS09oz?usp=sharing

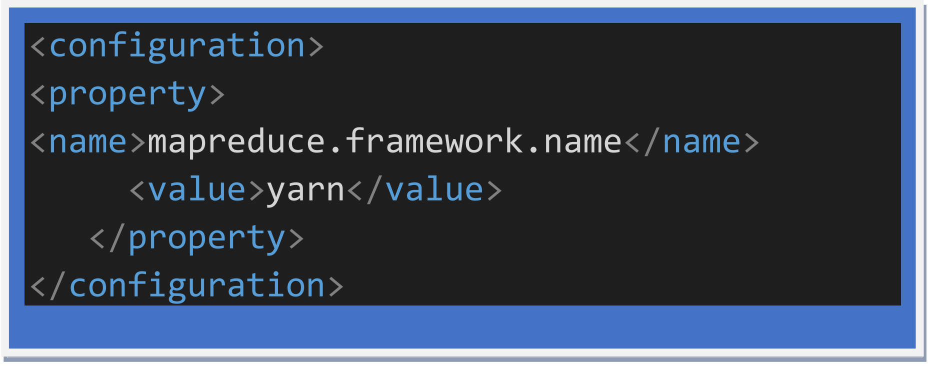
into the correct directory and modify them accordingly. If you do copy and paste from the PDF, open the files in Vscode and make sure they look EXACTLY like they do here, with NO extra characters.

C:\hadoop_bigdata\hadoop\etc\hadoop\core-site.xml



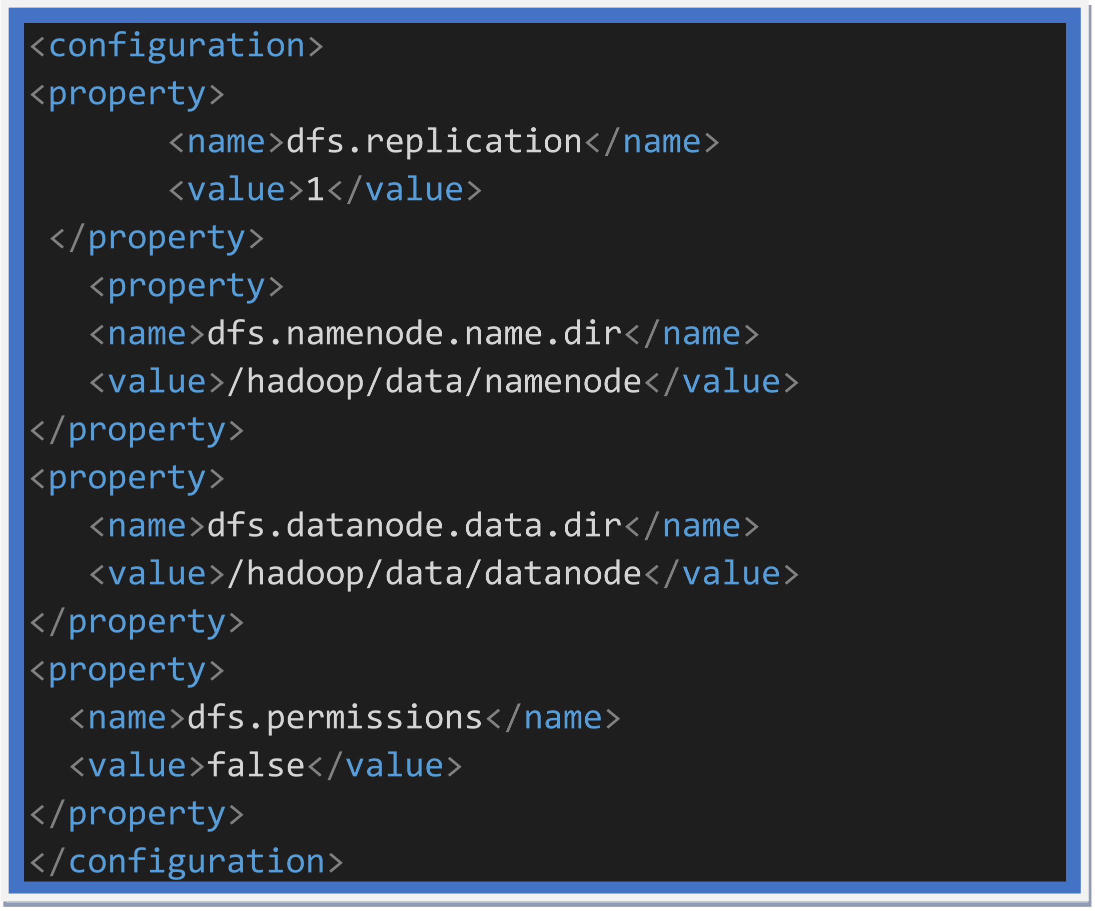
```xml
<configuration>
<property>
      <name>fs.defaultFS</name>
     <value>hdfs://localhost:9000</value>
   </property>
</configuration>
```

C:\hadoop_bigdata\hadoop\etc\hadoop\mapred-site.xml



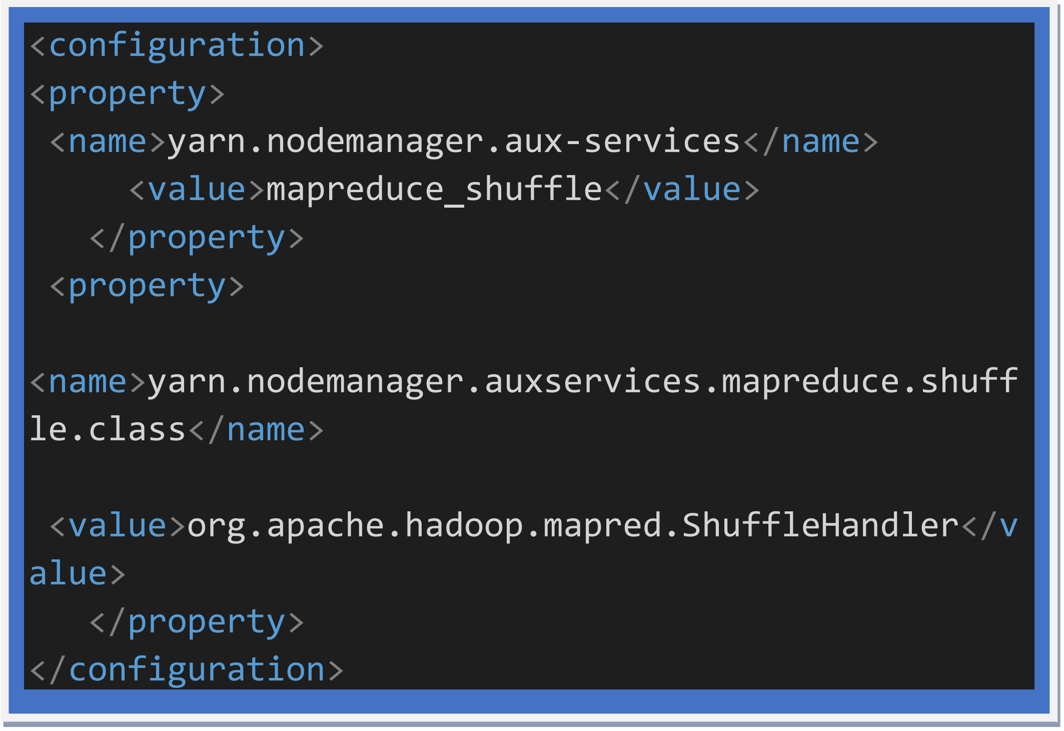
```xml
<configuration>
<property>
<name>mapreduce.framework.name</name>
     <value>yarn</value>
   </property>
</configuration>
```

C:\hadoop_bigdata\hadoop\etc\hadoop\hdfs-site.xml



```xml
<configuration>
<property>
       <name>dfs.replication</name>
       <value>1</value>
 </property>
   <property>
   <name>dfs.namenode.name.dir</name>
   <value>/hadoop/data/namenode</value>
</property>
<property>
   <name>dfs.datanode.data.dir</name>
   <value>/hadoop/data/datanode</value>
</property>
<property>
  <name>dfs.permissions</name>
  <value>false</value>
</property>
</configuration>
```

C:\hadoop_bigdata\hadoop\etc\hadoop\yarn-site.xml



```xml
<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>
```

- Now let's configure hadoop-env.cmd C:\hadoop_bigdata\hadoop\etc\hadoop\ open Hadoop-env.cmd with vscode or your fav IDE
- Search for set JAVA_HOME='Put your java location here ' in my case

```
set JAVA_HOME=C:\PROGRA~1\Java\JDK18~1.0_3
```

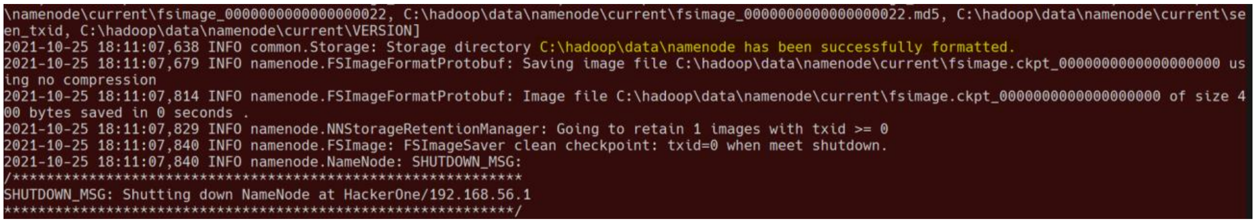
the Truth moment ❤

V. Start the Hadoop Cluster :

- Open up a Command Prompt (CMD)
- Format the NameNode:

```
C:\Users\abdou>hdfs namenode -format
```

You should then see the following output (note the **"C:\hadoop\data\namenode has been successfully formatted.** – that is what you want!)

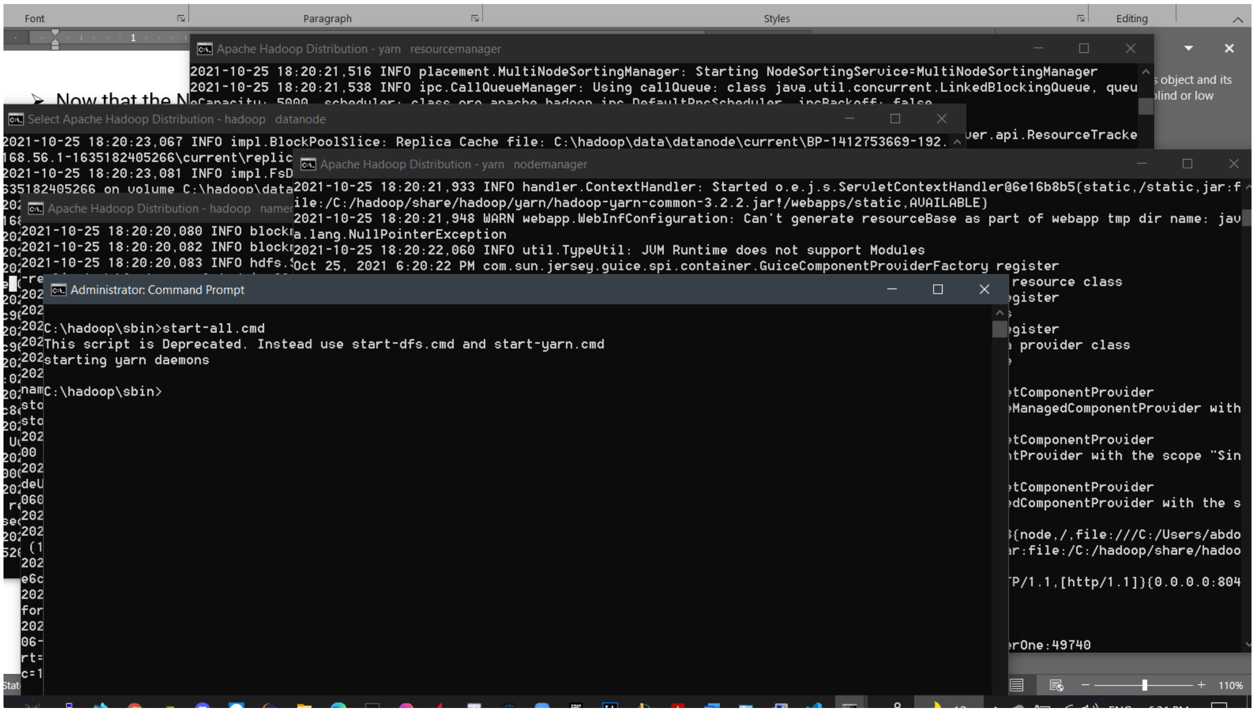


```
\namenode\current\fsimage_0000000000000000022, C:\hadoop\data\namenode\current\fsimage_0000000000000000022.md5, C:\hadoop\data\namenode\current\se
en_txid, C:\hadoop\data\namenode\current\VERSION]
2021-10-25 18:11:07,638 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2021-10-25 18:11:07,679 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_0000000000000000000 us
ing no compression
2021-10-25 18:11:07,814 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_0000000000000000000 of size 4
00 bytes saved in 0 seconds .
2021-10-25 18:11:07,829 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-10-25 18:11:07,840 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-10-25 18:11:07,840 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at HackerOne/192.168.56.1
************************************************************/
```

- Now that the NameNode is formatted, we have a usable HDFS directory. Let's start the HDFS dataemon now. **Allow any Windows Firewall prompts you see.**

- Finally locate to **C:\hadoop_bigdata\hadoop\sbin**
  And start CMD then type the following Command

```
C:\hadoop\sbin>start-all.cmd
```


Font
Paragraph
Styles
Editing
Apache Hadoop Distribution - yarn resourcemanager
Select Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn nodemanager
Apache Hadoop Distribution - hadoop namenode
Administrator: Command Prompt

```
2021-10-25 18:20:21,516 INFO placement.MultiNodeSortingManager: Starting NodeSortingService=MultiNodeSortingManager
2021-10-25 18:20:21,538 INFO ipc.CallQueueManager: Using callQueue: class java.util.concurrent.LinkedBlockingQueue, queu
```

```
2021-10-25 18:20:23,067 INFO impl.BlockPoolSlice: Replica Cache file: C:\hadoop\data\datanode\current\BP-1412753669-192.
168.56.1-1635182405266\current\replic
2021-10-25 18:20:23,081 INFO impl.FsD
635182405266 on volume C:\hadoop\data
```

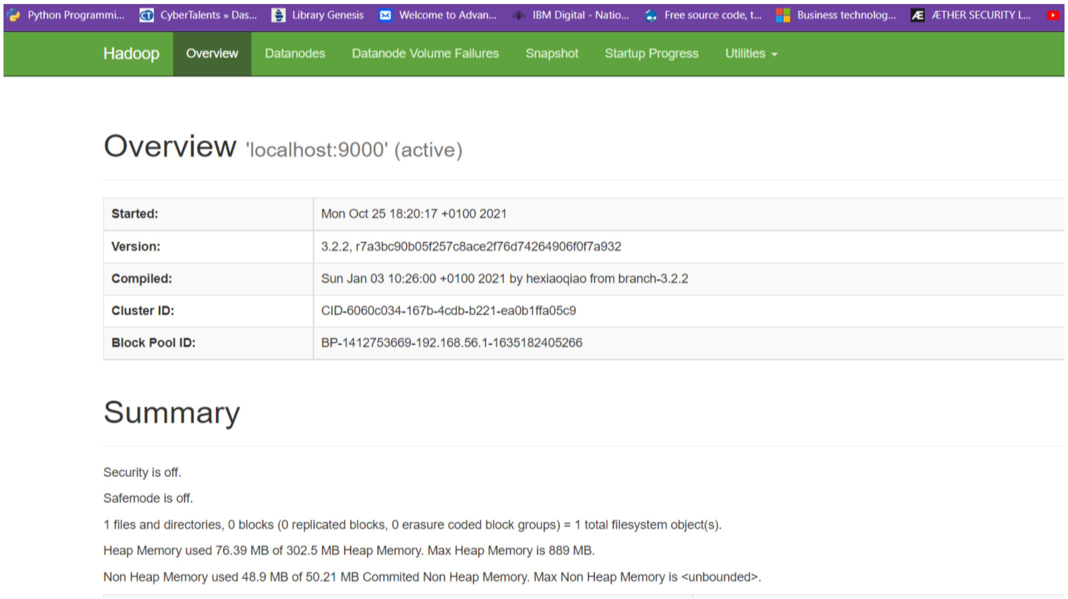
```
2021-10-25 18:20:20,080 INFO block
2021-10-25 18:20:20,082 INFO block
2021-10-25 18:20:20,083 INFO hdfs.
```

```
2021-10-25 18:20:21,933 INFO handler.ContextHandler: Started o.e.j.s.ServletContextHandler@6e16b8b5{static,/static,jar:f
ile:/C:/hadoop/share/hadoop/yarn/hadoop-yarn-common-3.2.2.jar!/webapps/static,AVAILABLE}
2021-10-25 18:20:21,948 WARN webapp.WebInfConfiguration: Can't generate resourceBase as part of webapp tmp dir name: jav
a.lang.NullPointerException
2021-10-25 18:20:22,060 INFO util.TypeUtil: JVM Runtime does not support Modules
Oct 25, 2021 6:20:22 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory register
```

```
C:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\hadoop\sbin>
```

- Launch your browser and type : http://localhost:9870/



Overview 'localhost:9000' (active)

Started:	Mon Oct 25 18:20:17 +0100 2021
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 10:26:00 +0100 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-6060c034-167b-4cdb-b221-ea0b1ffa05c9
Block Pool ID:	BP-1412753669-192.168.56.1-1635182405266

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 76.39 MB of 302.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 48.9 MB of 50.21 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

- If you made it this far on the first try, congrats! You now have a working Hadoop cluster ☺