

02 hadoop on windows

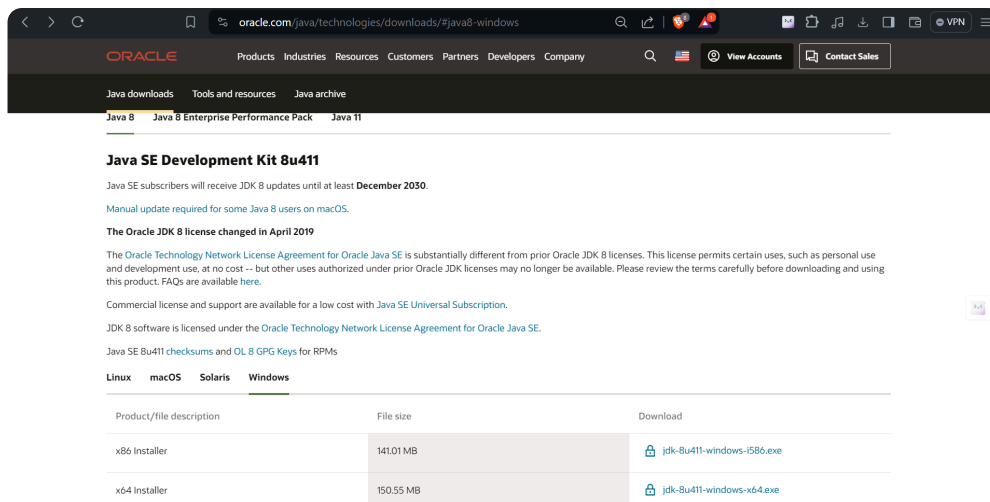
<https://youtu.be/knAS0w-jiUk?si=84sAh8kjZ46tUnO4>

1. Delete all Previous Java Versions
2. Create a folder in C drive "Java"

Install jdk 8

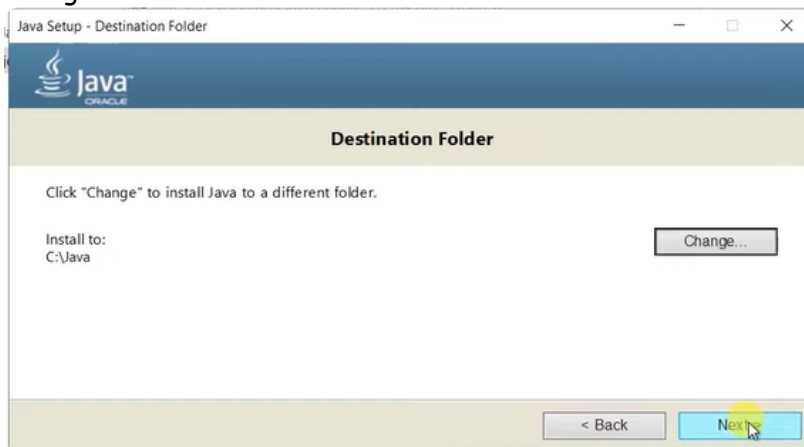
```
jdk-8u411-windows-x64.exe
```

<https://www.oracle.com/java/technologies/downloads/#java8-windows>

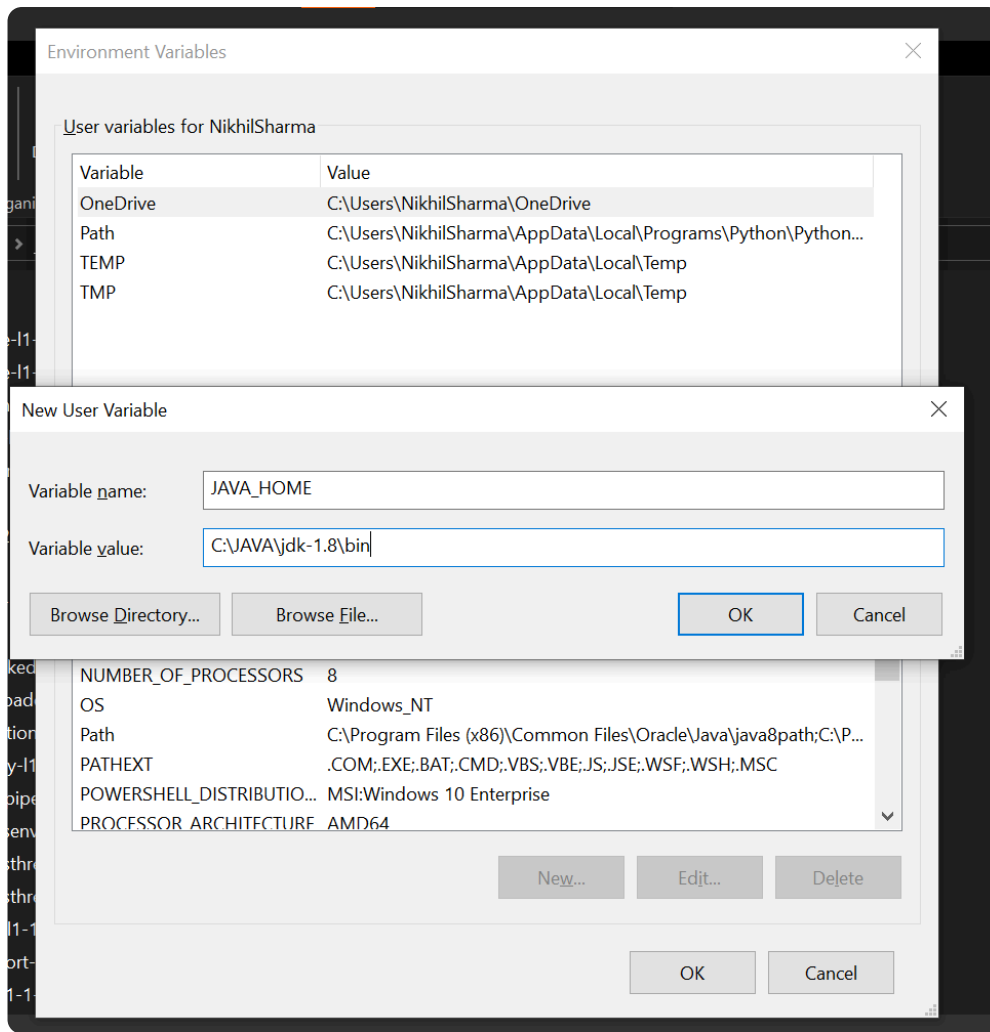


Run the installation, Prompt to change the folder while installing JDK, Choose the 'Java' folder in C:
When installation is done

- Change the folder the 2nd time when asked.



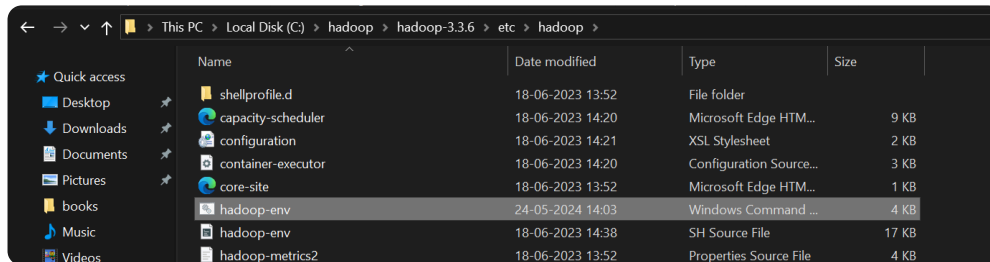
Go to C://Program Files and inside Java, cut the jdk1.8 folder put it inside the Java folder in C:



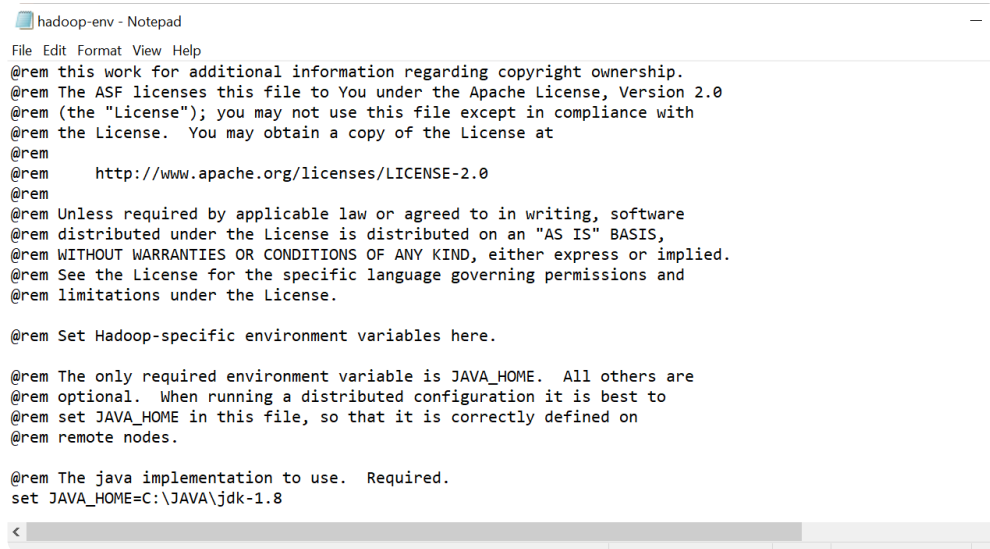
Also add the same path to the system variable path.

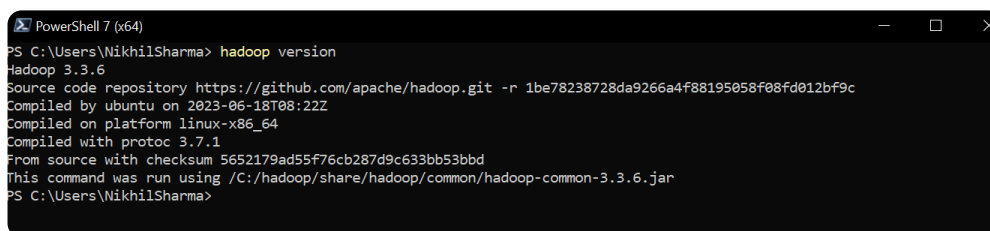
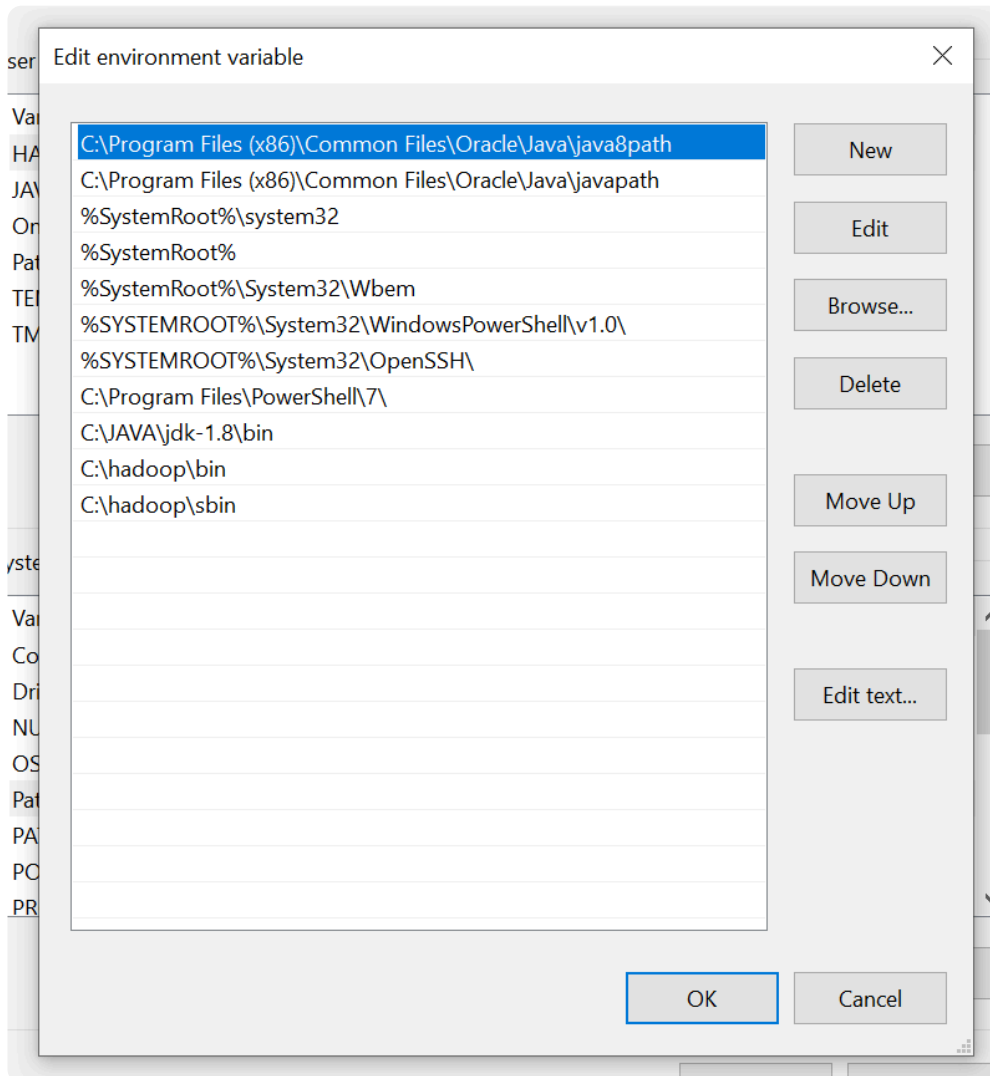
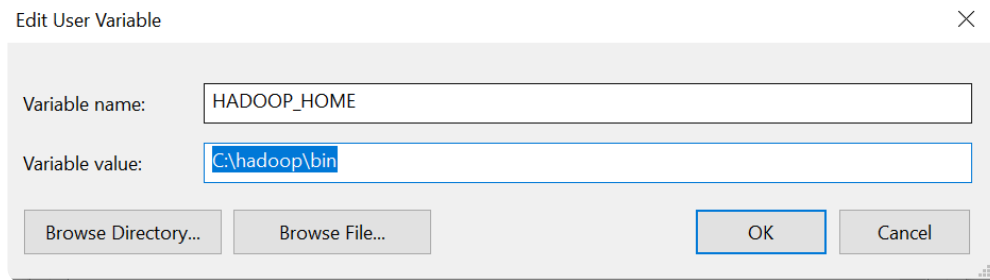
Download the binary file of the Apache Hadoop

<https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz>



Update the JAVA_HOME value in the hadoop env file





Edit the Hadoop Configuration Files

1. C:\hadoop\etc\hadoop
edit the core-site file

```
<configuration>

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

```
</configuration>
```

Create a folder "data" in C:\hadoop

This data folder should contain two more folders.

namenode

datanode

2. httpfs-site file

```
<configuration>

<property>
<name>dfs.replication</name>
<value>1</value>
</property>

<property>
<name>dfs.namenode.name.dir</name>
<value>C:\hadoop\data\namenode</value>
</property>

<property>
<name>dfs.datanode.data.dir</name>
<value>C:\hadoop\data\datanode</value>
</property>

</configuration>
```

3. mapred-site

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

4. yarn-site

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>

<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.shuffleHandler</value>
</property>

</configuration>
```

Fix the hadoop bin.

Delete the bin folder and download from [here](#)

and then Run the Winutils present in the bin folder.

One dll file might be missing, download it. MSVCR120.dll

and paste inside

```
C:\Windows\System32
```

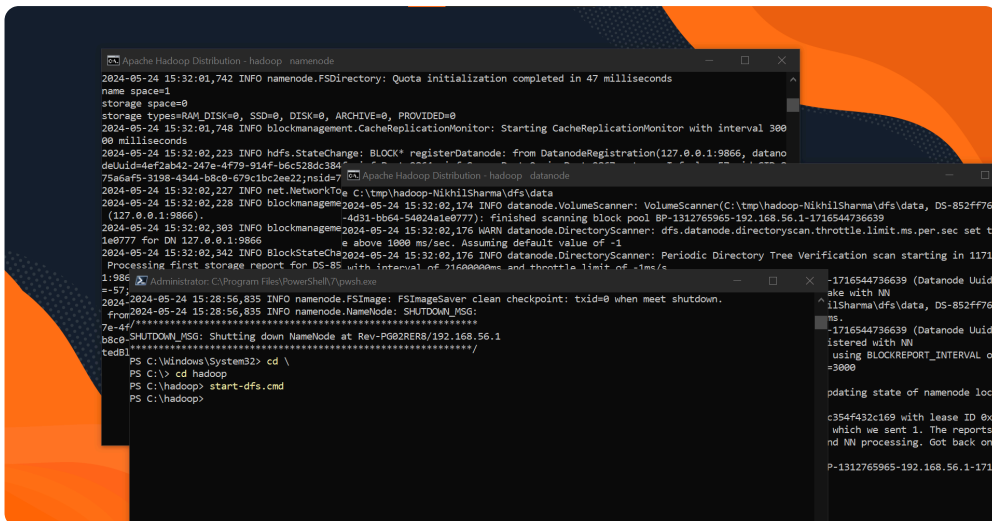
winutils inside the bin folder should not throw any error.

[Latest supported Visual C++ Redistributable downloads](#) | [Microsoft Learn](#)

run powershell as admin

```
hdfs namenode -format
```

```
cd \
cd hadoop
cd sbin
start-dfs.cmd
```



The `jps` command in Hadoop stands for "Java Process Status."

- check the status of Java processes running on a Hadoop cluster.
- lists the Java processes along with their process IDs (PIDs).

```
PS C:\hadoop> jps
10576 DataNode
5020 Jps
9740 NameNode
```

Run yarn

```
start-yarn.cmd
```

```
PS C:\hadoop> jps
10576 DataNode
6976 Jps
18200 NodeManager
16460 ResourceManager
9740 NameNode
PS C:\hadoop>
```

Overview 'localhost:9000' (✓active)

Started:	Fri May 24 15:32:00 +0530 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f06fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-875a6af5-3198-4344-b8c0-679c1bc2ee22
Block Pool ID:	BP-1312765965-192.168.56.1-1716544736639

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: 122.19 MB of 343.5 MB Heap Memory. Max Heap Memory is 889 MB.

To see Resource Manager
localhost:8088

Cluster

- About
- Nodes
- Node Labels
- Applications
 - NEW
 - NEW_SAVING
 - SUBMITTED
 - ACCEPTED
 - RUNNING
 - FINISHED
 - FAILED
 - KILLED
- Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	User
0	0	0	0	0	<memory:0 B, vC

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	Fin
Showing 0 to 0 of 0 entries											

stop-all.cmd to stop all