02 hadoop on windows

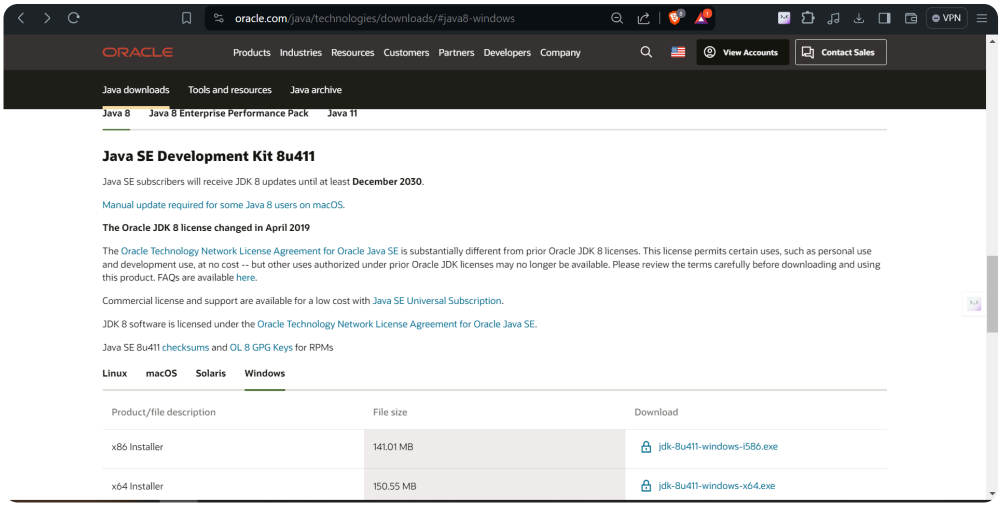
https://youtu.be/knAS0w-jiUk?si=84sAh8kjZ46tUnO4

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

Install jdk 8

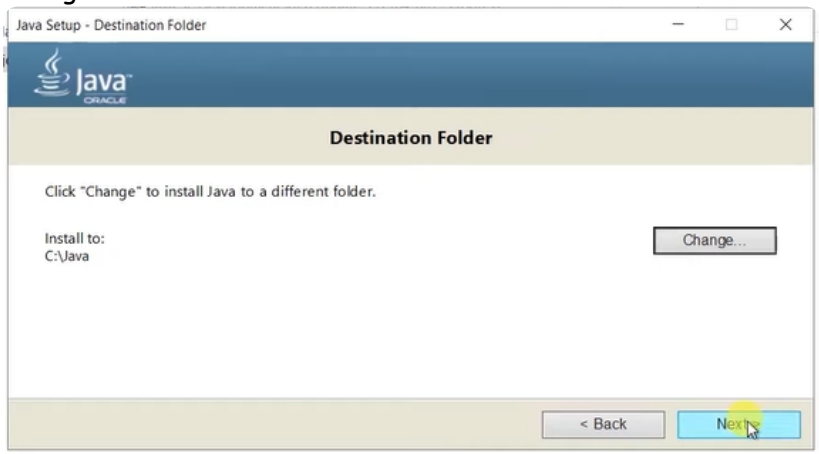
```
jdk-8u411-windows-x64.exe
```

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

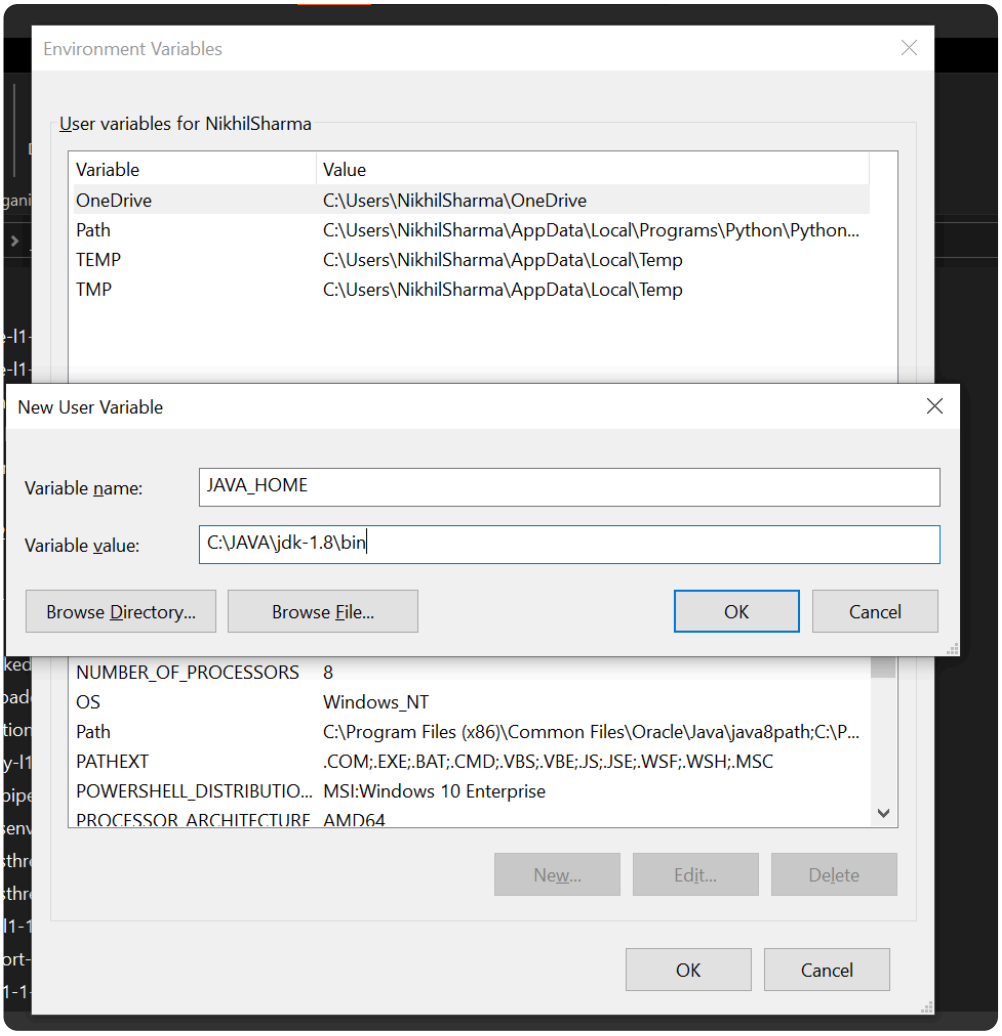


Run the installation, Prompt to change the folder while installing JDK, Choose the 'Java' folder in C:
When installation is done

- Change the folder the 2nd time when asked.



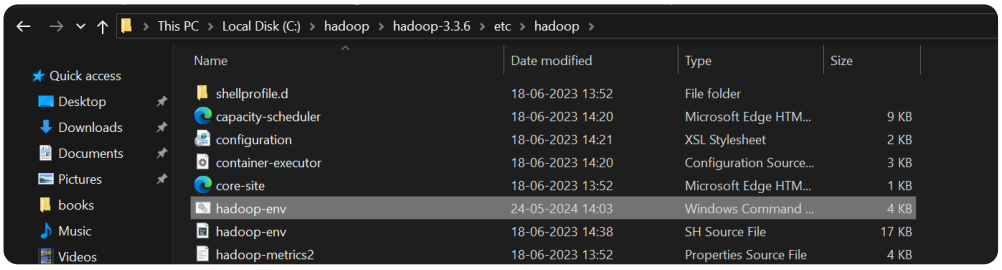
Go to C://Program Files and inside Java, cut the jdk1.8 folder put it inside the Java folder in C:



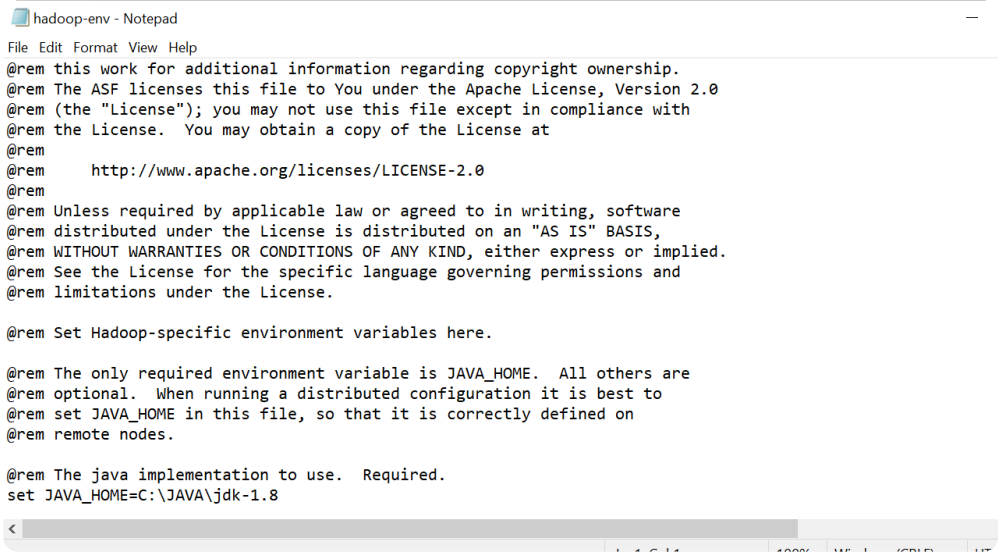
Also add the same path to the system variable path.

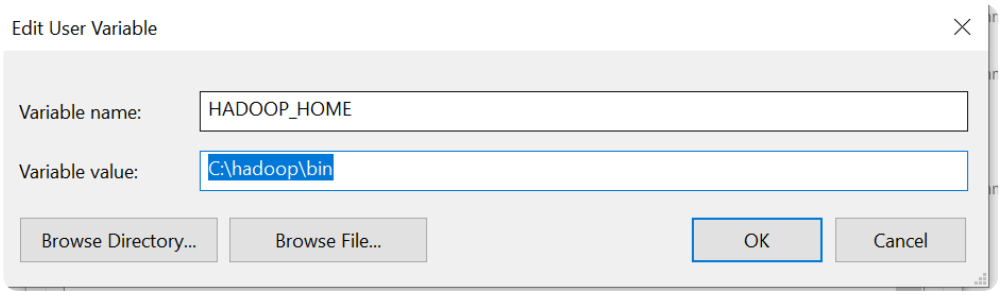
Download the binary file of the Apache Hadoop

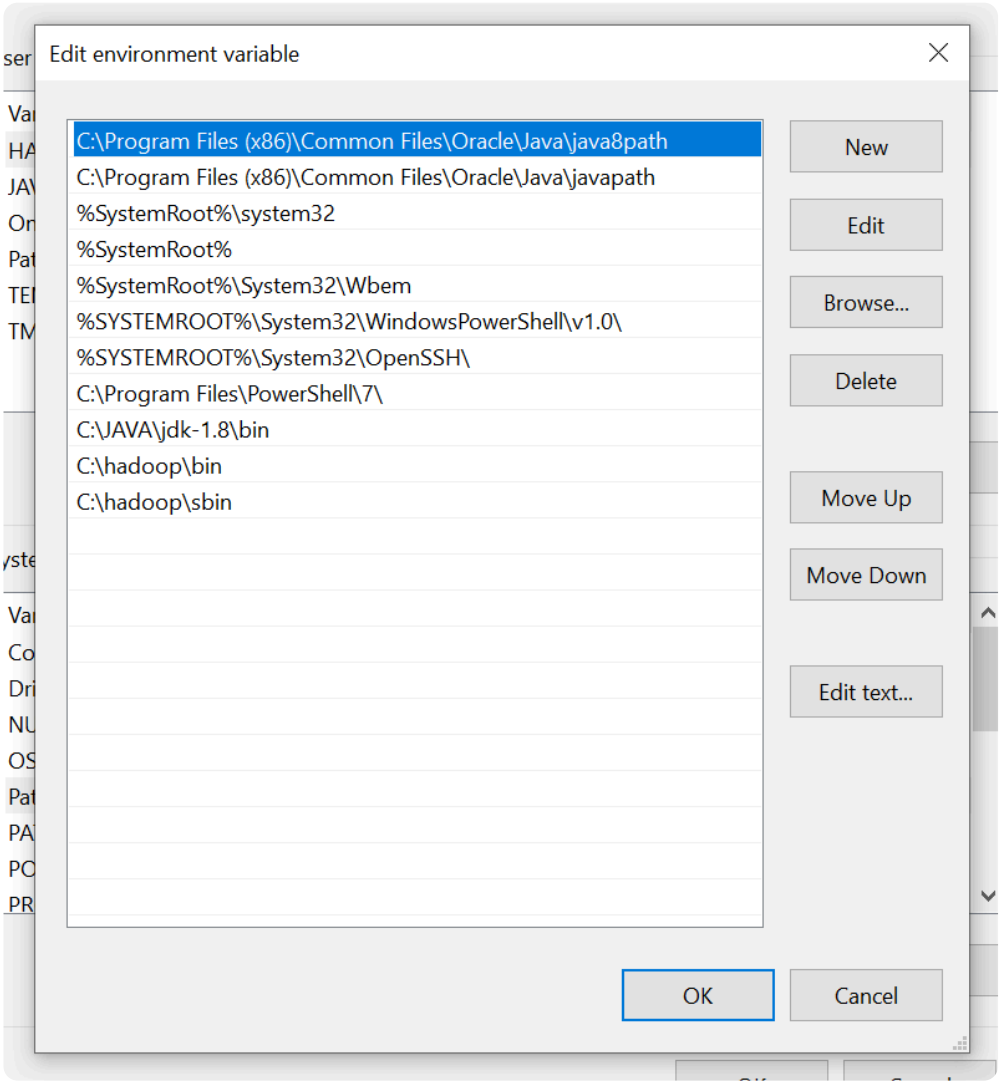
https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz



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>
```

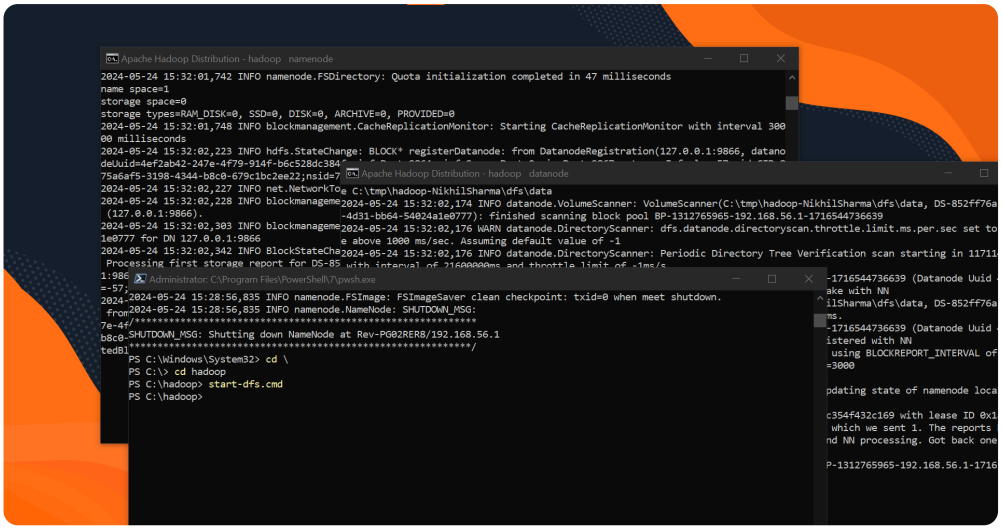
4. `yarn-site`

```
<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
```



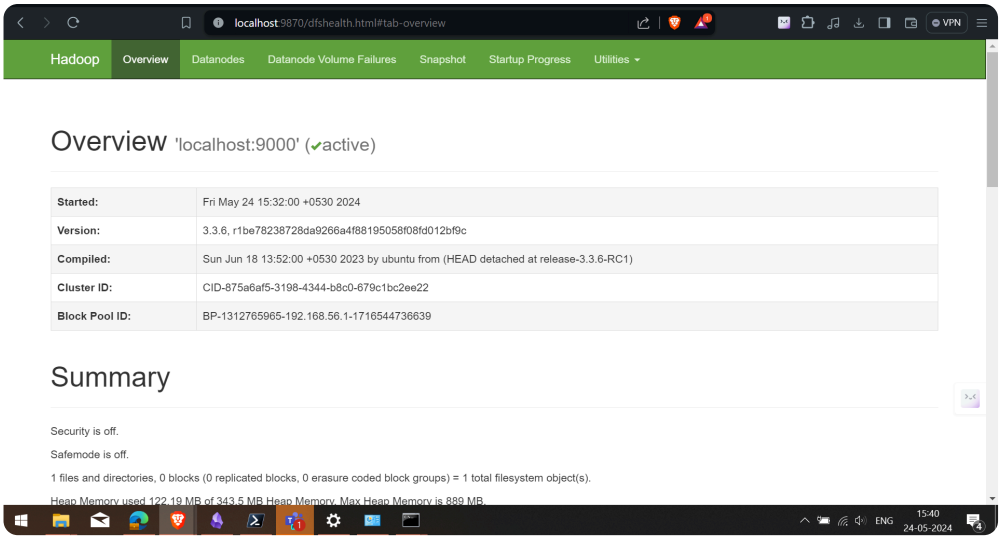
The `jps` command in Hadoop stands for "Java Process Status."

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

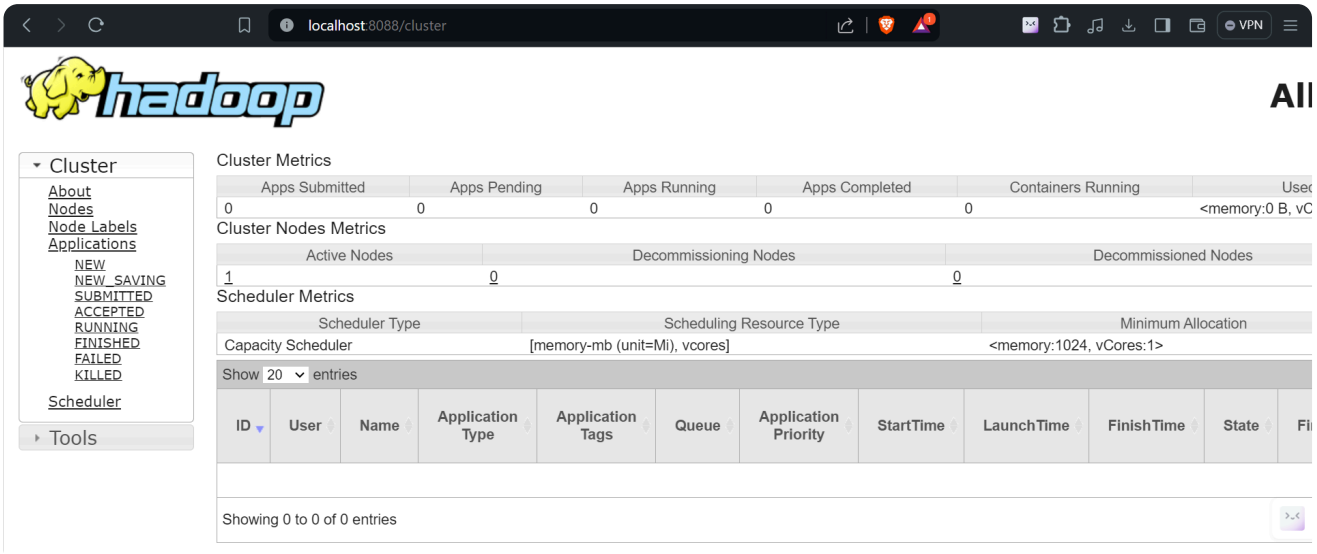
```
PS C:\hadoop> jps
10576 DataNode
5020 Jps
9740 NameNode
```

Run yarn
`start-yarn.cmd`

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



To see Resource Manager
localhost:8088



`stop-all.cmd` to stop all