

# Hadoop Installation

in Windows

# Prerequisites

1. Install java JDK
2. Install WSL from powershell. (wsl --install).
3. Download latest hadoop(tar.gz) from [apache official website](#).
4. Install wget if possible.

**Note: Do not install java into the default program files directory instead install it into a directory whose name contains no space in any of its path**

# Hadoop without Docker

1. Extract the downloaded hadoop tar.gz file

use this command : `tar -xvf hadoop_file.tar.gz`

2. Create a folder named Hadoop and copy the extracted tar file into it

3. If possible create a folder named jdkxx in Hadoop and install java jdk into it.

4. Setup JAVA\_HOME and HADOOP\_HOME variables in environment system variables.

# Environment variables setup

1. Go to settings and search for “Environment Variables” and select “Edit the system environment variables”.
2. In Advanced click on Environment Variables.
3. Under System variables, click on New then
  - a. Variable name: JAVA\_HOME
  - b. Variable value: Path to jdk installation directory
    - i.example : In my case it is D:\Hadoop\jdk17
4. Under System variables, click on New then
  - a. Variable name: HADOOP\_HOME
  - b. Variable value: Path to hadoop directory where you extracted tar file
    - i.example : In my case D:\Hadoop\hadoop-3.3.6
5. Edit the Path variable and add bin directories of jdk and hadoop to it
  - a.example : In my case it is D:\Hadoop\{**jdk17** | **hadoop-3.3.6**}\bin

# Verifying the Installation

1. Open command prompt in hadoop home directory
2. Execute command “hadoop version”. If installation is correct you need to see the output as below

```
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/Users/MahaaGURU/Downloads/Hadoop/hadoop-3.3.6/share/hadoop/common/hadoop-common-3.3.6.jar
```

# Setup namenode and datanode directories

1. To setup namenode and datanode directories do the following below steps
  - a. Create folder named name and data.
    - i. example : I created in “C” drive and modified the permissions of the file for the hadoop to have full control of the directory.
    - ii. namenode path : C:\tmp\hadoop-user\dfs\name
    - iii. datanode path : C:\tmp\hadoop-user\dfs\data
2. The namenode maintains the critical information such as HDFS namespace, block locations and the file system tree.
3. The datanode maintains the actual data blocks of HDFS files

# Setup namenode and datanode directories

1. Now open “{HADOOP\_HOME}\etc\hadoop\hdfs-site.xml” and add the following lines in it.

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:///C:/tmp/hadoop-MahaaGURU/dfs/name</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:///C:/tmp/hadoop-MahaaGURU/dfs/data</value>
</property>
</configuration>
```

Note: “dfs.replication” property says number of copies of each blocks should be made. Ideally it should be 3 in production environment, for testing purpose 1 is fine.



# Setup core-site.xml

1. Now open “{HADOOP\_HOME}\etc\hadoop\core-site.xml” and add the following lines in it.

```
<configuration>
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://localhost:9000</value>
</property>
<property>
  <name>hadoop.tmp.dir</name>
  <value>C:/tmp/hadoop-MahaaGURU</value>
  <description>A base for other temporary directories.</description>
</property>
</configuration>
```

Note: “fs.defaultFs” indicates the URI of the namenode service and “hadoop.tmp.dir” indicates the temporary directory which is used to save the temporary data while running mapreduce program.



# Start HDFS

1. Open command prompt in admin mode and navigate to hadoop home directory
2. Execute the command “hadoop namenode -format”
3. Execute the command “sbin\start-dfs.cmd”.
4. It takes some time to start namenode and datanode. Once done run command “jps”, you should see the status as below in cmd.
5. To stop hdfs run the command sbin\stop-dfs.cmd.

```
C:\Users\MahaaGURU\Downloads\Hadoop\hadoop-3.3.6>jps
24336 DataNode
21988 NameNode
24892 Jps
```

Thank You