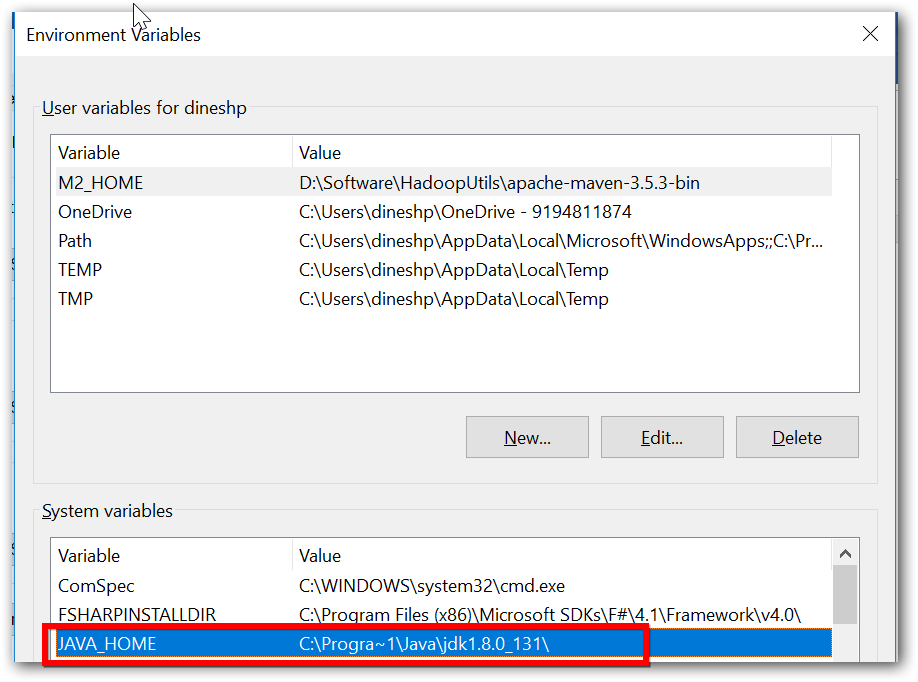
Hadoop + Azure Data Lake Gen 1

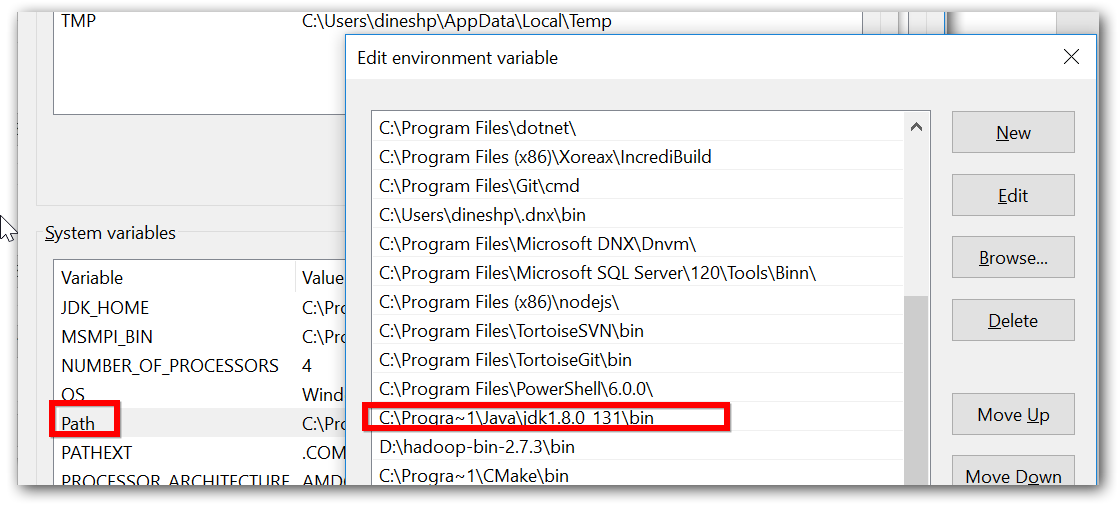
# Pre-requisites

1. Install **7z1801-x64** in your machine.
2. Untar **“hadoop-3.0.1.tar”** using 7z. Prefer any other drive apart from OS drive. For e.g. if ‘C:\’ is OS drive, have Hadoop package in ‘D:\’ or some other drive.
3. Install Java 8 in your machine.
4. Set JAVA\_HOME in environment variables – system variables

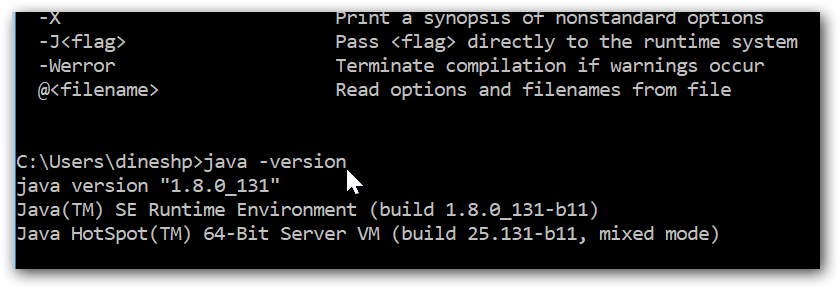
<https://confluence.atlassian.com/doc/setting-the-java_home-variable-in-windows-8895.html>



1. Add Java’s bin location PATH

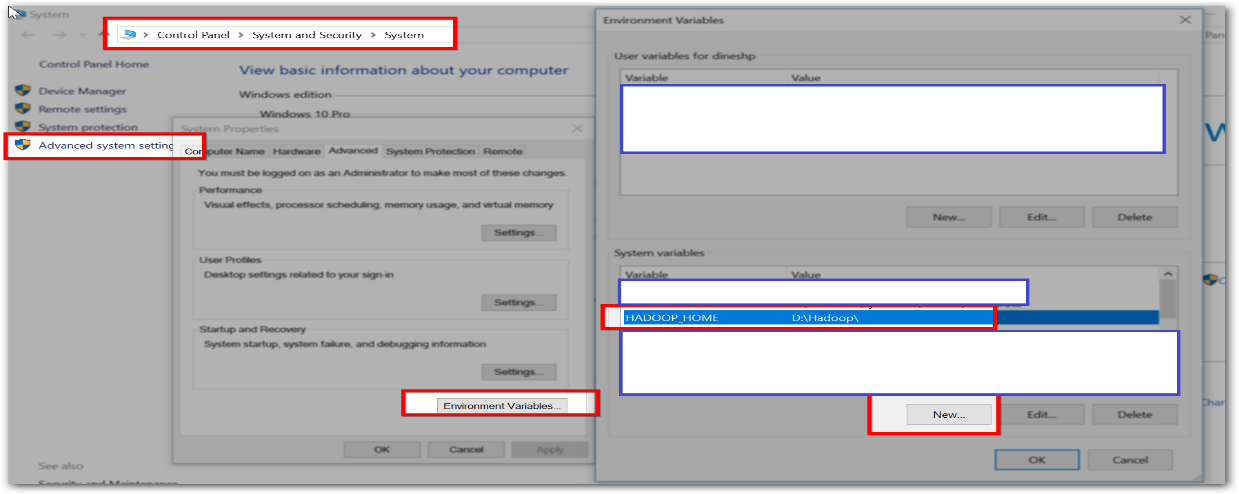


1. Ensure java home and path is set properly.
   1. Open command prompt
   2. Execute ‘javac’ or ‘java -version’ command.



1. Set HADOOP\_HOME as environment variables (without bin),

|  |  |
| --- | --- |
| HADOOP\_HOME | D:\hadoop\3.0.1 |



# Hadoop cluster installation

## Configuration

1. For Azure Data Lake Gen 1 storage you need to copy the following jar files to “D:\hadoop-3.0.1\share\hadoop\hdfs\lib”
   1. azure-storage-8.0.0.jarhadoop-azure-datalake-3.0.0-alpha2.jar
   2. azure-data-lake-store-sdk-2.0.11.jar
2. Do the following changes in specified file name located in directory “D:\hadoop-3.0.1\etc\hadoop\”

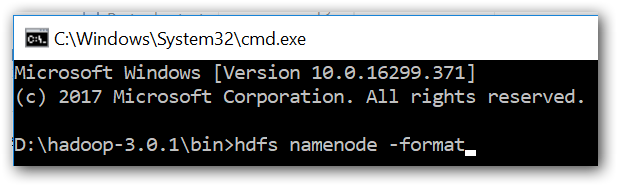
|  |  |  |
| --- | --- | --- |
| **File name** | **Properties** | **Comments** |
| hadoop-env.cmd  (Windows command script file) | set **HADOOP\_PREFIX**=D:\hadoop-3.0.1  set HADOOP\_CONF\_DIR=%HADOOP\_PREFIX%\etc\hadoop  set YARN\_CONF\_DIR=%HADOOP\_CONF\_DIR%  set PATH=%PATH%;%HADOOP\_PREFIX%\bin | Add at the *end* of file, after 90th line.  **HADOOP\_PREFIX** is the Hadoop package unzipped location. |
| core-site.xml | <configuration>  <property>  <name>fs.defaultFS</name>  <value>hdfs://localhost:9000</value>  </property>  <property>  <name>fs.adl.oauth2.access.token.provider.type</name>  <value>ClientCredential</value>  </property>  <property>  <name>fs.adl.oauth2.refresh.url</name>  <value>**TOKEN ENDPOINT**</value>  </property>  <property>  <name>fs.adl.oauth2.client.id</name>  <value>**APPLICATION ID**</value>  </property>  <property>  <name>fs.adl.oauth2.credential</name>  <value>**CLIENT SECRET**</value>  </property>  <property>  <name>fs.adl.impl</name>  <value>org.apache.hadoop.fs.adl.AdlFileSystem</value>  </property>  <property>  <name>fs.AbstractFileSystem.adl.impl</name>  <value>org.apache.hadoop.fs.adl.Adl</value>  </property>  </configuration> | Replace the content instead of <configuration></configuration>  Changes in **bold** are required if you need to configure data lake gen1 storage in the cluster. If not specified, the cluster will be with local storage alone. |
| hdfs-site.xml | <configuration>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  <property>  <name>dfs.namenode.name.dir</name>  <value>file:///d:/Data/NameNode</value>  </property>  <property>  <name>dfs.datanode.data.dir</name>  <value>file:///d:/Data/DataNode</value>  </property>  </configuration> | Replace the content instead of <configuration></configuration>  Prefer having any other drive apart from ‘C:\’ |

## Format your cluster and make your file system ready for use

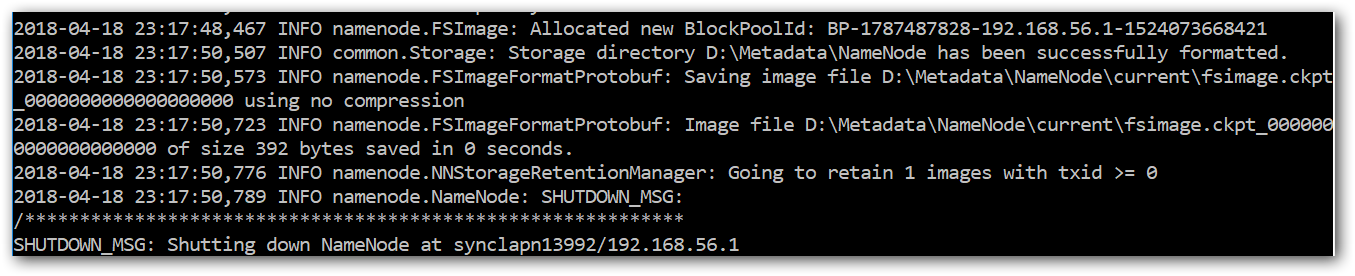
Open command prompt as ‘Administrator’ and navigate to Hadoop bin path,

1. Format the cluster

\bin> hdfs namenode -format



1. Ensure that format got successfully completed with similar message in below image,

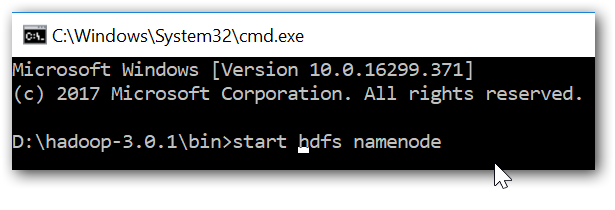


## Start Hadoop cluster services

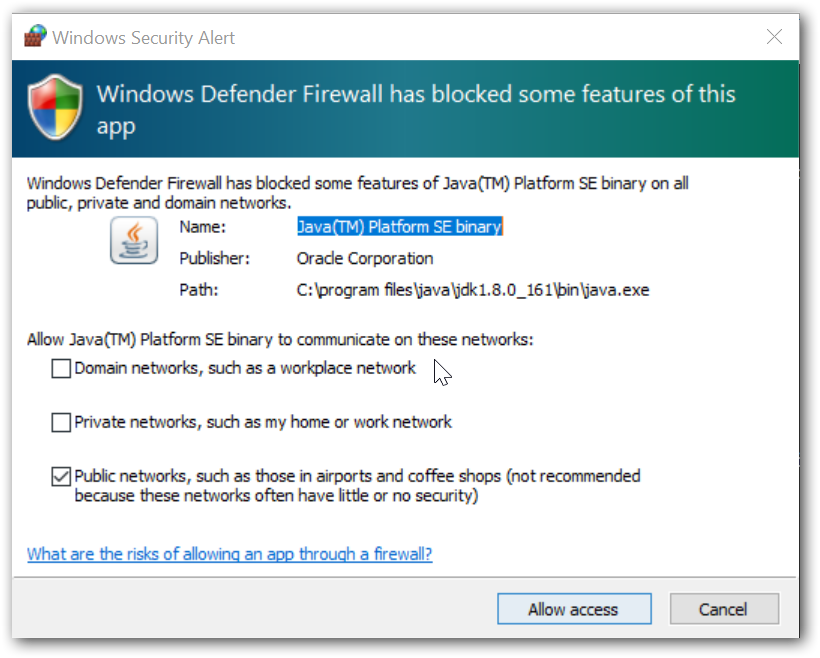
Open command prompt as ‘Administrator’ and navigate to Hadoop bin path,

1. Start NameNode

\bin> start hdfs namenode

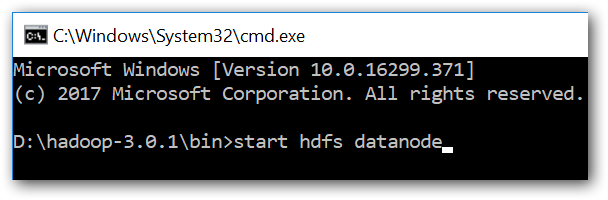


If such a popup is showed while starting service, click ‘Allow access’



1. Start DataNode

\bin> start hdfs datanode



1. Web UI - HDFS - <http://localhost:9870/dfshealth.html#tab-overview>

# Work-out samples

Open command prompt as ‘Administrator’ and navigate to Hadoop bin path,

1. List the directories in Azure Data Lake store

hdfs dfs -ls adl://datalakedatasource.azuredatalakestore.net