**Hadoop and Hive Setup on Windows 🖥️**

**🌟 Key Highlights**

* **Hadoop Setup:** Configuring HDFS and running distributed computations.
* **Hive Setup:** Installing Hive and integrating it with Hadoop.
* **Troubleshooting:** Solutions to common issues faced during the setup.

**🛠️ Prerequisites**

* **Software Requirements:**
  + [Java JDK 8](https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html)
  + [Hadoop 3.2.4](https://hadoop.apache.org/release/3.2.4.html)
  + [Apache Hive 3.1.2](https://archive.apache.org/dist/hive/hive-3.1.2/)

**Important:** Please ensure to download these exact specified versions from the provided links to avoid version compatibility issues. These versions have been tested and work well together.

**📁 Directory Structure**

* README.md: Main documentation.
* config-files: Configuration files like core-site.xml and hive-site.xml.

**🚀 Getting Started**

Follow the detailed steps in this README to complete your setup successfully.

**☕ Java Installation**

**Step 1: Download Java JDK**

1. Visit the [Java JDK 8](https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html).
2. Select the appropriate version for your Windows system:
   * **x64 Installer** for 64-bit Windows.
   * **x86 Installer** for 32-bit Windows.
3. Download the .exe file and run the installer.

**Step 2: Install Java**

1. Follow the installer prompts:
   * **Choose Installation Directory**: Instead of the default directory, create a new folder in the C: drive named Java (e.g., C:\Java).
   * Avoid installing in Program Files to prevent issues caused by spaces in the path.
   * Complete the installation process.
2. After installation, note the installation path (e.g., C:\Java\jdk-<version>).

**Step 3: Set Environment Variables**

1. Open **Control Panel** → **System** → **Advanced System Settings** → **Environment Variables**.
2. Under **User Variables**, click **New**:
   * **Variable Name**: JAVA\_HOME
   * **Variable Value**: Path to your JDK folder (e.g., C:\Java\jdk-<version>).
3. Edit the Path variable:
   * Add %JAVA\_HOME%\bin to the list.

Note: Replace the path with the actual path to jdk folder in your system

**Step 4: Verify Installation**

1. Open a Command Prompt (Win + R, then type cmd).
2. Run the command:
3. java -version

javac -version

**Expected Outputs**

**Command: java -version**

java version "1.8.0\_431"

Java(TM) SE Runtime Environment (build 1.8.0\_431-b10)

Java HotSpot(TM) 64-Bit Server VM (build 25.431-b10, mixed mode)

**Command: java -version**

javac 1.8.0\_431

**🛠️ Troubleshooting**

If java -version does not work:

* Ensure the JAVA\_HOME path is correct.
* Ensure %JAVA\_HOME%\bin is added to the Path variable.

**🐘 Hadoop Installation**

**Step 1: Download Hadoop**

1. Visit [Hadoop 3.2.4](https://hadoop.apache.org/release/3.2.4.html).
2. Download the specified version of Hadoop (i.e., Hadoop 3.2.4).
3. Extract the downloaded .tar.gz file into a directory, such as C:\Hadoop.

**Important:** When extracting the .tar.gz file, ensure the path does not have spaces. Extract the files into a directory like C:\Hadoop and **rename the folder** if it is something like hadoop-3.2.4 to just Hadoop for consistency.

**Step 2: Set Environment Variables**

1. Open **Control Panel** → **System** → **Advanced System Settings** → **Environment Variables**.
2. Under **User Variables**, click **New**:
   * **Variable Name**: HADOOP\_HOME
   * **Variable Value**: Path to your hadoop bin folder (i.e., C:\hadoop).
3. Edit the Path variable:
   * Add C:\hadoop\bin to the list.
   * Add C:\hadoop\sbin to the list.

**Step 3: Configure Hadoop**

1. **Edit hadoop-env.cmd:**
   * Navigate to C:\Hadoop\etc\hadoop\.
   * Open hadoop-env.cmd in a text editor.
   * Set the JAVA\_HOME variable to your Java installation path:

set JAVA\_HOME=C:\Java\jdk-<version>

Note: Replace the path with the actual path to jdk folder in your system

1. **Configure core-site.xml:**
   * Open C:\Hadoop\etc\hadoop\core-site.xml and add the following configuration:
   * <configuration>
   * <property>
   * <name>fs.defaultFS</name>
   * <value>hdfs://localhost:9000</value>
   * </property>

</configuration>

1. **Configure hdfs-site.xml:**
   * Before editing the hdfs-site.xml file, **create a new folder** called data in C:\Hadoop, and inside this data folder, create **two subfolders** named namenode and datanode.

Note: Ensure all folder names are in small letters and make sure the spelling is correct.

* + Now, open C:\Hadoop\etc\hadoop\hdfs-site.xml and add the following configuration:
  + <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>

1. **Configure mapred-site.xml:**
   * Open C:\Hadoop\etc\hadoop\mapred-site.xml and add the following configuration:
   * <configuration>
   * <property>
   * <name>mapreduce.framework.name</name>
   * <value>yarn</value>
   * </property>

</configuration>

1. **Configure yarn-site.xml:**
   * Open C:\Hadoop\etc\hadoop\yarn-site.xml and add the following configuration:
   * <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>

1. **Download and Replace bin Folder**
   * Download the bin zip file from this [Google Drive link](https://drive.google.com/drive/folders/1iURNbow2IglhAhSy3sfY5xxVfAg33NBW).
   * Extract the bin folder from the downloaded zip file.
   * Navigate to C:\Hadoop.
   * Delete the existing bin folder.
   * Paste the newly downloaded bin folder into the C:\Hadoop directory, replacing the old one.

**Step 4: Start Hadoop Services**

1. **Run Command Prompt as Administrator:**
   * Right-click on the Command Prompt and select **Run as Administrator**.
2. **Navigate to the sbin directory:**
   * Open Command Prompt as Administrator.
   * Navigate to the sbin directory:

cd C:\Hadoop\sbin

1. **Stop Hadoop Services:**
   * Stop the Hadoop Distributed File System (HDFS) and YARN:
   * stop-dfs.cmd

stop-yarn.cmd

1. **Format the DataNode and NameNode:**
   * Format the DataNode:

hdfs datanode -format

* + Format the NameNode:

hdfs namenode -format

1. **Start Hadoop Services:**
   * Start the Hadoop Distributed File System (HDFS) and YARN:
   * start-dfs.cmd

start-yarn.cmd

**Step 5: Verify Hadoop Services with jps Command**

1. **Verify if Hadoop services are running:**
   * After starting the Hadoop services, run the jps command to check the status of the running processes:

jps

1. **Expected Output:**
   * The output should look similar to this:
   * 21728 Jps
   * 13524 NameNode
   * 16520 DataNode
   * 17180 NodeManager
   * 5628 ResourceManager
   * If you see these processes listed, it means Hadoop services are running correctly.

**Step 6: Access Hadoop Dashboards**

1. **NameNode Web UI:**
   * Open your browser and visit:
   * http://localhost:9870
   * This is the **NameNode Web UI**, where you can monitor the HDFS.
2. **Hadoop Cluster Resource Manager:**
   * Open your browser and visit:
   * http://localhost:8088/cluster
   * This is the **Hadoop Cluster Resource Manager**, where you can monitor YARN applications and resources.

**🐝 Hive Installation**

**1. Download and Configure Apache Derby**

a) Download Derby 10.14.2.0 (compatible with Hadoop 3.2.4) from [Apache Derby Downloads](https://db.apache.org/derby/derby_downloads.html).  
b) Extract the file using WinRar, 7Zip, or similar software.  
c) Copy the extracted folder to C:\ and rename it to derby (e.g., C:\derby).

**2. Download and Configure Apache Hive Binaries**

a) Download Hive 3.1.2 (compatible with Hadoop 3.2.4) from [Apache Hive Archives](https://archive.apache.org/dist/hive/hive-3.1.2/).  
b) Extract the file using WinRar, 7Zip, or similar software.  
c) Copy the extracted folder to C:\ and rename it to hive (e.g., C:\hive).  
d) Navigate to C:\derby\lib and copy all the \*.jar files.  
e) Paste the copied \*.jar files into C:\hive\lib.  
f) Download the exact hive-site.xml file from my [Google Drive](https://drive.google.com/file/d/1_h8dN46KK8zZf9yf2M2DBUfPR_IZl33T/view).  
g) Copy the downloaded hive-site.xml file and paste it into the C:\hive\conf folder.

**3. Cross-Check Guava JAR File**

a) Ensure C:\hive\lib contains the same version of the guava-x.y-jre.jar file as the Hadoop version located in C:\hadoop\share\hadoop\common\lib.

* For example: If Hadoop uses guava-27.0-jre.jar, Hive must also use guava-27.0-jre.jar.  
  b) If there is a mismatch, copy the guava-27.0-jre.jar from Hadoop and paste it into C:\hive\lib, then delete any other versions of the guava-x.y-jre.jar file.

**4. Create Temporary Folder**

a) Create a folder named tmp in C:\hive (e.g., C:\hive\tmp).  
b) Set the folder properties to grant all rights.

**5. Set Up Environmental Variables**

1. Add the following variables:
   * DERBY\_HOME: C:\derby
   * HIVE\_HOME: C:\hive
   * HIVE\_BIN: C:\hive\bin
   * HIVE\_LIB: C:\hive\lib
   * HADOOP\_USER\_CLASSPATH\_FIRST: true
2. Add the following paths to the system Path variable:
   * C:\derby\bin
   * C:\hive\bin

**6. Download and Configure Wget**

a) Download wget.exe from [Eternally Bored](https://eternallybored.org/misc/wget/1.21.4/64/wget.exe).  
b) Copy wget.exe to C:\wget.  
c) Add C:\wget to the system Path variable.

**7. Download Windows Version of Hive Bin**

a) Navigate to a directory and create a folder of your choice (e.g., C:\bin).  
b) Open Command Prompt and run:

wget -r -np -nH --cut-dirs=3 -R index.html https://svn.apache.org/repos/asf/hive/trunk/bin/

c) Navigate to C:\bin and its subdirectories until you find the bin folder. d) Copy the bin folder. e) Navigate to C:\hive, delete the existing bin folder, and replace it with the downloaded bin folder.

**8. Run Hadoop Daemons**

a) Start Hadoop daemons:

start-all.cmd

b) Check if daemons are running:

jps

**9. Run Derby Server**

a) Start the Derby Network Server:

StartNetworkServer -h 0.0.0.0

**10. Initialize Hive Schema**

a) Navigate to C:\hive\bin and run the following command:

hive --service schematool -dbType derby -initSchema

**11. Start Hive**

a) Start Hive by running:

hive