

Hadoop 3.1.2 Installation on Windows

Pre-requisite software

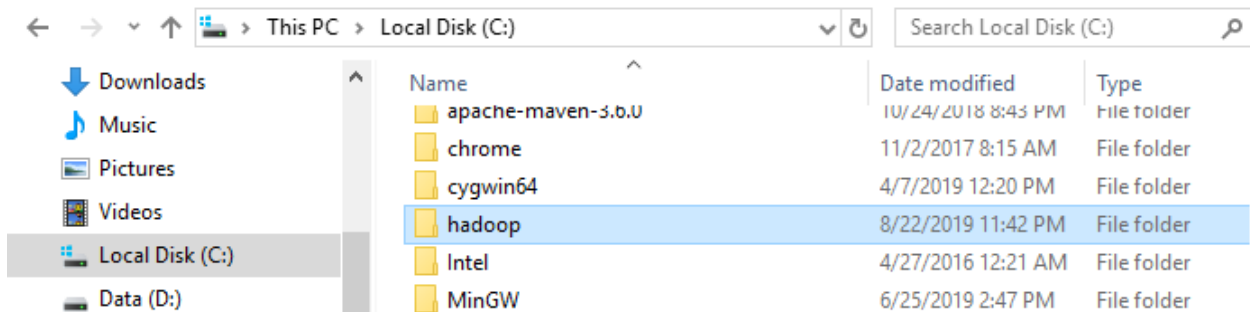
Java version 1.8 (JDK)

```
javac -version
```

Download Hadoop 3.1.2 binaries

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz>

Pada tutorial ini, direktori binaries Hadoop disimpan di drive C. Copy file yang telah didownload ke drive C lalu extract dan ganti nama foldernya dengan “hadoop”.



Download Windows Compatible Binaries

bit.do/winutils

Folder bin yang terdapat pada folder hadoop yang baru saja diextract hanya compatible untuk linux. Maka dari itu, harus diganti oleh binaries yang compatible dengan OS Windows.

Download folder bin dari link di atas, extract, lalu copy ke direktori hadoop (folder bin yang lama direname saja). Contohnya dapat dilihat pada gambar berikut:

| Local Disk (C:) > hadoop | | | |
|--------------------------|--------------------|-------------|------|
| Name | Date modified | Type | Size |
| bin | 8/22/2019 11:31 PM | File folder | |
| bin.old | 8/22/2019 11:28 PM | File folder | |

Create folder for datanode and namenode

Buat folder sebagai berikut:

- Folder datanode di C:/hadoop/data/datanode
- Folder namenode di C:/hadoop/data/namenode

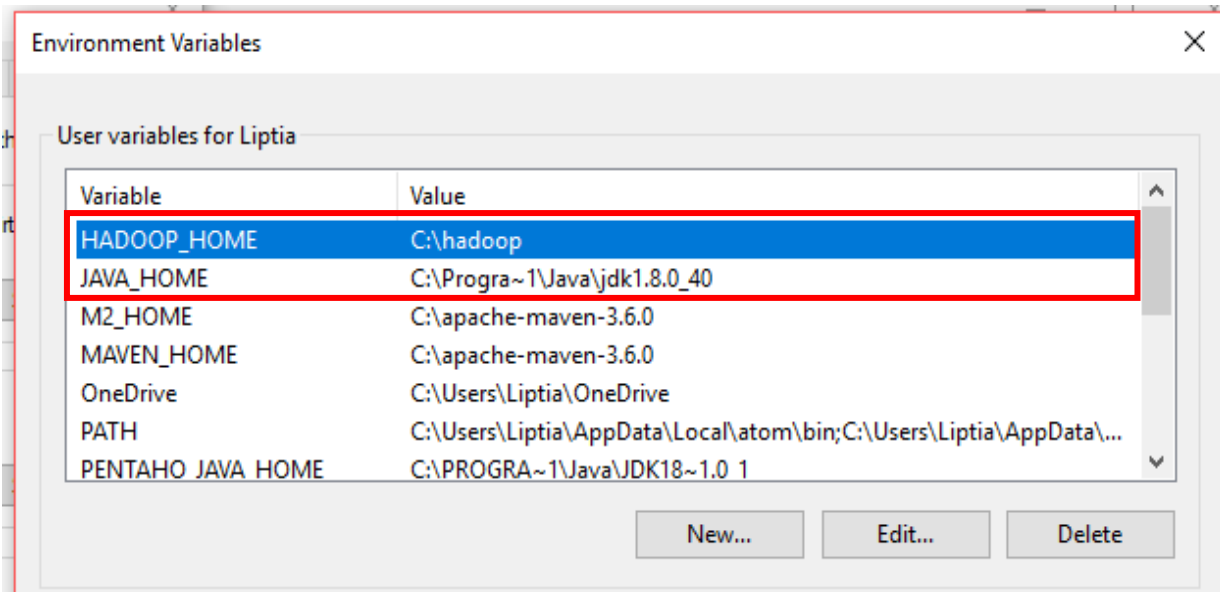
| Local Disk (C:) > hadoop | | | |
|--------------------------|--------------------|---------------|--------|
| Name | Date modified | Type | Size |
| bin | 8/22/2019 11:31 PM | File folder | |
| bin.old | 8/22/2019 11:28 PM | File folder | |
| data | 8/22/2019 11:33 PM | File folder | |
| etc | 8/22/2019 11:28 PM | File folder | |
| include | 8/22/2019 11:31 PM | File folder | |
| lib | 8/22/2019 11:28 PM | File folder | |
| libexec | 8/22/2019 11:28 PM | File folder | |
| logs | 8/22/2019 11:43 PM | File folder | |
| sbin | 8/22/2019 11:28 PM | File folder | |
| share | 8/22/2019 11:28 PM | File folder | |
| LICENSE.txt | 1/23/2019 10:07 PM | Text Document | 144 KB |
| NOTICE.txt | 1/23/2019 10:07 PM | Text Document | 22 KB |
| README.txt | 1/23/2019 10:07 PM | Text Document | 2 KB |

Set Hadoop environment variables

Hadoop membutuhkan dua environment variable yang perlu diset.

- HADOOP_HOME="C:\hadoop"
- JAVA_HOME=<Path di mana JDK diinstall>

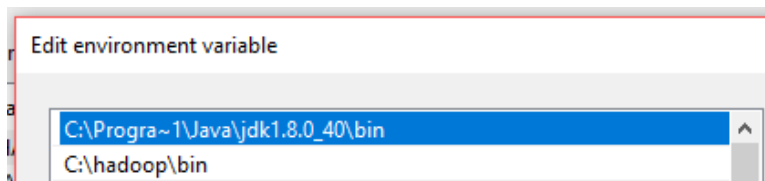
Windows+break -> Advanced System settings -> Environment variables. Click New untuk membuat environment variable baru.



Edit Path environment variable

Path variable pada system variables diedit lalu tambahkan 2 baris berikut satu per satu:

- C:\hadoop\bin
- <Path di mana JDK diinstall>\bin



Setelah selesai, untuk memeriksa apakah hadoop sudah terinstall dengan benar, pada command prompt ketik:

```
hadoop version
```

```
C:\WINDOWS\system32>hadoop version
Hadoop 3.1.2
Source code repository https://github.com/apache/hadoop.git -r 1019dde65bcf12e05ef48ac71e84550d589e5d9a
Compiled by sunilg on 2019-01-29T01:39Z
Compiled with protoc 2.5.0
From source with checksum 64b8bdd4ca6e77cce75a93eb09ab2a9
This command was run using /C:/hadoop/share/hadoop/common/hadoop-common-3.1.2.jar
```

Configure Hadoop

Terdapat 5 file konfigurasi yang harus diedit untuk mengkonfigurasi Hadoop:

1. **hadoop-env.cmd**
2. **core-site.xml**
3. **hdfs-site.xml**
4. **mapred-site.xml**
5. **yarn-site.xml**

1. Edit hadoop-env.cmd

Buka file C:\hadoop\etc\hadoop\hadoop-env.cmd lalu edit path JAVA_HOME seperti berikut ini:

```
set JAVA_HOME=<Path di mana JDK diinstall>
```

```
23
24 @rem The java implementation to use.  Required.
25 @rem set JAVA_HOME=%JAVA_HOME%
26 set JAVA_HOME=C:\Progra~1\Java\jdk1.8.0_40
27
```

Di bagian paling bawah file tersebut, tambahkan beberapa baris berikut:

```
set HADOOP_CLASSPATH=%JAVA_HOME%\lib\tools.jar

set HADOOP_PREFIX=C:\hadoop

set HADOOP_CONF_DIR=C:\hadoop\etc\hadoop

set YARN_CONF_DIR=C:\hadoop\etc\hadoop

set PATH=%PATH%;C:\hadoop\bin
```

```
93
94 set HADOOP_CLASSPATH=%JAVA_HOME%\lib\tools.jar
95 set HADOOP_PREFIX=C:\hadoop
96 set HADOOP_CONF_DIR=C:\hadoop\etc\hadoop
97 set YARN_CONF_DIR=C:\hadoop\etc\hadoop
98 set PATH=%PATH%;C:\hadoop\bin
99
```

2. Edit core-site.xml

Edit file C:\hadoop\etc\hadoop**core-site.xml**, kemudian tambahkan properti berikut di dalam tag configuration:

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value> hdfs://localhost:50071</value>
  </property>
</configuration>
```

3. Edit hdfs-site.xml

Selanjutnya mengkonfigurasi jumlah replikasi data block dan lokasi folder datanode serta namenode yang sudah dibuat sebelumnya. Edit file C:\hadoop\etc\hadoop**hdfs-site.xml** lalu tambahkan beberapa properti berikut:

```
<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>
```

4. Edit mapred-site.xml

Edit file C:\hadoop\etc\hadoop\mapred-site.xml untuk mengedit konfigurasi dari framework Map-Reduce. Jika file tersebut tidak ada, copy file **mapred-site.xml.template** lalu ubah nama filenya menjadi **mapred-site.xml**

```
<configuration>

    <property>

        <name>mapreduce.framework.name</name>

        <value>yarn</value>

    </property>

    <property>

        <name>mapred.job.tracker</name>

        <value>localhost:9001</value>

    </property>

</configuration>
```

5. Edit yarn-site.xml

Edit file C:\hadoop\etc\hadoop\yarn-site.xml dengan menambahkan beberapa property sebagai berikut:

```
<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>

  <property>
    <name>yarn.nodemanager.delete.debug-delay-sec</name>
    <value>600</value>
  </property>

  <property>
    <name>yarn.application.classpath</name>
    <value>%HADOOP_HOME%\etc\hadoop,
%HADOOP_HOME%\share\hadoop\common\*, %HADOOP_HOME%\share\hadoop\common\lib\*,
%HADOOP_HOME%\share\hadoop\mapreduce\*, %HADOOP_HOME%\share\hadoop\mapreduce\lib\*,
%HADOOP_HOME%\share\hadoop\hdfs\*, %HADOOP_HOME%\share\hadoop\hdfs\lib\*,
%HADOOP_HOME%\share\hadoop\yarn\*, %HADOOP_HOME%\share\hadoop\yarn\lib\*</value>
  </property>
</configuration>
```

Edit file Slaves

Pada direktori **C:\hadoop\etc\hadoop** pastikan ada file bernama **slaves**. Jika tidak ada, maka buat lalu tambahkan **localhost** di dalam file tersebut.

| | | | |
|-------------------------------|--------------------|--------------------|------|
| mapred-env.cmd | 1/29/2019 9:58 AM | Windows Comma... | 1 KB |
| mapred-env.sh | 1/29/2019 9:58 AM | SH File | 2 KB |
| mapred-queues.xml.template | 1/29/2019 9:58 AM | Dev-C++ Templat... | 5 KB |
| mapred-site.xml | 8/22/2019 11:35 PM | XML Document | 1 KB |
| slaves | 8/22/2019 6:55 AM | File | 1 KB |
| ssl-client.xml.example | 1/29/2019 8:41 AM | EXAMPLE File | 3 KB |
| ssl-server.xml.example | 1/29/2019 8:41 AM | EXAMPLE File | 3 KB |
| user_ec_policies.xml.template | 1/29/2019 8:52 AM | Dev-C++ Templat... | 3 KB |
| ... | 1/29/2019 8:41 AM | FILE | 1 KB |

Format Namenode

Buka Command Prompt dan run sebagai administrator. Lalu jalankan perintah berikut:

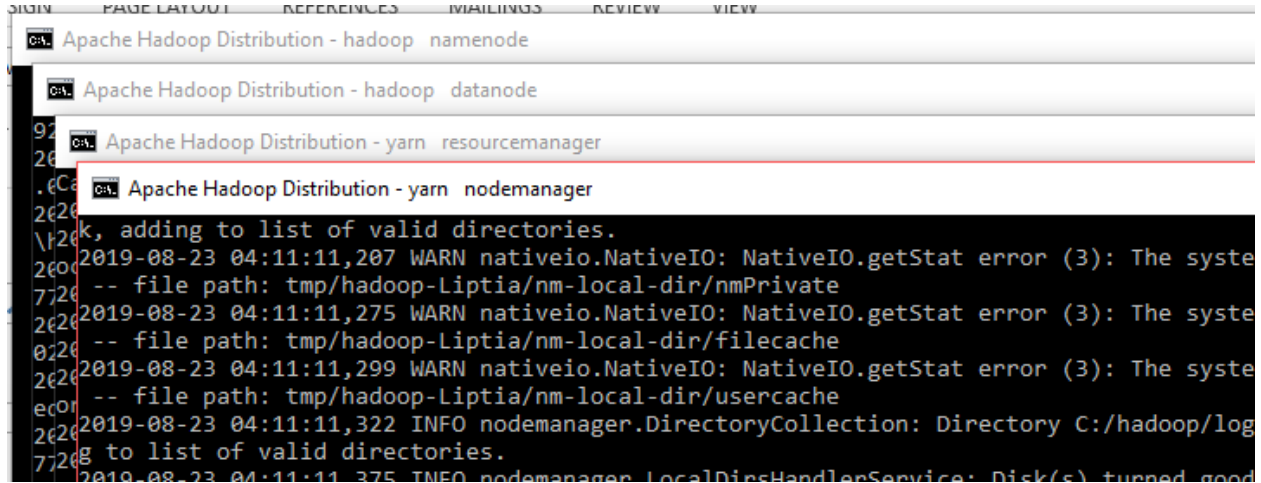
```
hdfs namenode -format
```

Launch Hadoop

Pada direktori **C:\hadoop\sbin**, run as administrator file **start-all.cmd**.

| Local Disk (C:) > hadoop > sbin | | | | |
|---------------------------------|--------------------|------------------|------|--|
| Name | Date modified | Type | Size | |
| FederationStateStore | 8/22/2019 11:28 PM | File folder | | |
| distribute-exclude.sh | 1/29/2019 8:52 AM | SH File | 3 KB | |
| hadoop-daemon.sh | 1/29/2019 8:41 AM | SH File | 2 KB | |
| hadoop-daemons.sh | 1/29/2019 8:41 AM | SH File | 3 KB | |
| https.sh | 1/29/2019 8:55 AM | SH File | 2 KB | |
| kms.sh | 1/29/2019 8:44 AM | SH File | 2 KB | |
| mr-jobhistory-daemon.sh | 1/29/2019 9:58 AM | SH File | 2 KB | |
| refresh-namenodes.sh | 1/29/2019 8:52 AM | SH File | 3 KB | |
| start-all.cmd | 1/29/2019 8:41 AM | Windows Comma... | 2 KB | |
| start-all.sh | 1/29/2019 8:41 AM | SH File | 3 KB | |
| start-balancer.sh | 1/29/2019 8:52 AM | SH File | 2 KB | |

Pastikan 4 windows cmd baru terbuka, yang memperlihatkan process dari 4 daemons hadoop



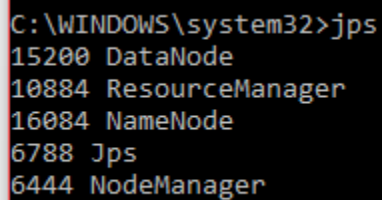
The screenshot shows four overlapping Windows Command Prompt windows. The top window is titled 'Apache Hadoop Distribution - hadoop namenode'. The second window is titled 'Apache Hadoop Distribution - hadoop datanode'. The third window is titled 'Apache Hadoop Distribution - yarn resourcemanager'. The bottom window is titled 'Apache Hadoop Distribution - yarn nodemanager' and displays the following log output:

```
2019-08-23 04:11:11,207 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system
-- file path: tmp/hadoop-Liptia/nm-local-dir/nmPrivate
2019-08-23 04:11:11,275 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system
-- file path: tmp/hadoop-Liptia/nm-local-dir/filecache
2019-08-23 04:11:11,299 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system
-- file path: tmp/hadoop-Liptia/nm-local-dir/usercache
2019-08-23 04:11:11,322 INFO nodemanager.DirectoryCollection: Directory C:/hadoop/log
g to list of valid directories.
2019-08-23 04:11:11,375 INFO nodemanager.LocalDirsHandlerService: Disk(s) turned good
```

Check Hadoop Daemons

Buka Command Prompt dan run sebagai administrator. Lalu jalankan perintah berikut:

```
jps
```



The screenshot shows a Windows Command Prompt window with the following output:

```
C:\WINDOWS\system32>jps
15200 DataNode
10884 ResourceManager
16084 NameNode
6788 Jps
6444 NodeManager
```

Hadoop Web UI

Resource manager -> localhost:8088

The screenshot shows the Hadoop Resource Manager Web UI. The top navigation bar includes the Hadoop logo and the title 'All Applications'. The left sidebar contains a 'Cluster' menu with options like 'About Nodes', 'Node Labels', 'Applications', and 'Scheduler'. The main content area displays several metrics tables:

- Cluster Metrics:** A table with columns for Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, Memory Total, Memory Reserved, VCores Used, VCores Total, and VCores Reserved. All values are 0.
- Cluster Nodes Metrics:** A table with columns for Active Nodes, Decommissioning Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, Rebooted Nodes, and Shutdown Nodes. All values are 0.
- Scheduler Metrics:** A table with columns for Scheduler Type, Scheduling Resource Type, Minimum Allocation, Maximum Allocation, and Maximum Cluster Application Priority. The values are Capacity Scheduler, [memory-mb (unit=Mi), vcores], <memory:1024, vCores:1>, <memory:8192, vCores:4>, and 0.

Below these metrics is a table of applications. The table is currently empty, showing 'No data available in table'. The table has columns for ID, User, Name, Application Type, Queue, Application Priority, StartTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU VCores, Allocated Memory MB, Reserved CPU VCores, Reserved Memory MB, % of Queue, % of Cluster, Progress, Tracking UI, and Blacklisted Nodes.

Namenode & datanode -> localhost:9870

The screenshot shows the Hadoop Namenode & Datanode Web UI. The top navigation bar includes the Hadoop logo and the title 'Overview'. The left sidebar contains a 'Cluster' menu with options like 'About Nodes', 'Node Labels', 'Applications', and 'Scheduler'. The main content area displays the 'Overview' section for 'localhost:50071' (active). The overview section includes a table of metadata:

| Property | Value |
|----------------|--|
| Started: | Thu Aug 22 23:43:03 +0700 2019 |
| Version: | 3.1.2, r1019dde65bcf12e05ef48ac71e84550d589e5d9a |
| Compiled: | Tue Jan 29 08:39:00 +0700 2019 by sunilg from branch-3.1.2 |
| Cluster ID: | CID-a0ae429d-284e-4493-9a5b-174456a4f54e |
| Block Pool ID: | BP-194252958-127.0.0.1-1566492141262 |

Below the metadata table is a 'Summary' section.

Stop Hadoop

Pada direktori C:\hadoop\sbin, run as administrator file **stop-all.cmd**.

The screenshot shows a Windows File Explorer window with the address bar set to 'Local Disk (C:) > hadoop > sbin'. The search bar contains 'Search sbin'. The file list shows the following files:

| Name | Date modified | Type |
|--------------------|-------------------|-------------|
| start-yarn.sh | 1/29/2019 9:53 AM | SH File |
| stop-all.cmd | 1/29/2019 8:41 AM | Windows Cor |
| stop-all.sh | 1/29/2019 8:41 AM | SH File |
| stop-balancer.sh | 1/29/2019 8:52 AM | SH File |
| stop-dfs.cmd | 1/29/2019 8:52 AM | Windows Cor |
| stop-dfs.sh | 1/29/2019 8:52 AM | SH File |
| stop-secure-dns.sh | 1/29/2019 8:52 AM | SH File |