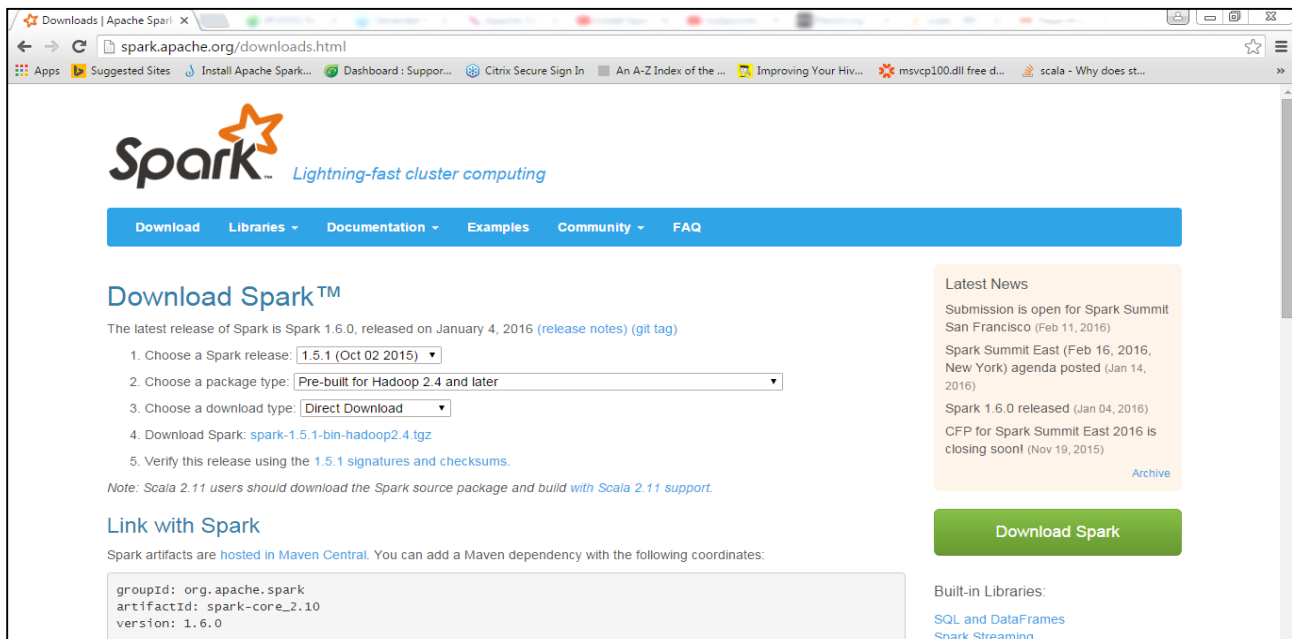


## SPARK – INSTALLATION GUIDE (WINDOWS 7)

**Step 1:** Go to the following link and download Spark:

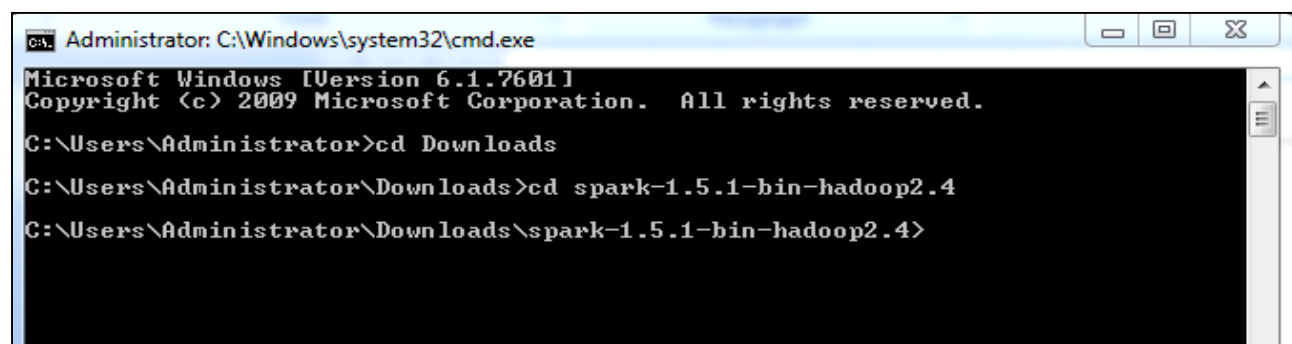
- <http://spark.apache.org/downloads.html>



**Step 2:** After downloading, extract the .jar file into a particular folder where you want to keep Spark.

**Step 3:** After extracting the .jar file, use the following commands to start Spark:

- Change the directory using `cd` command:



- bin\spark-shell

```
C:\Users\Administrator> C:\Windows\system32\cmd.exe - bin\spark-shell
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4> bin\spark-shell
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.lib.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.properties
To adjust logging level use sc.setLogLevel("INFO")
Welcome to

  ____  __
 / ___/ /  _  \
/ /   / _/  _/
/ /___/_/  _/
/____/_/  _/

version 1.5.1

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_79)
Type in expressions to have them evaluated.
Type :help for more information.
16/02/16 17:00:03 WARN MetricsSystem: Using default name DAGScheduler for source
because spark.app.id is not set.
Spark context available as sc.
16/02/16 17:00:06 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
istered. Ensure you dont have multiple JAR versions of the same plugin in the cl
asspath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1-bin-hadoop2
.4/bin/./lib/datanucleus-core-3.2.10.jar" is already registered, and you are tr
ying to register an identical plugin located at URL "file:/C:/Users/Administrato
r/Downloads/spark-1.5.1-bin-hadoop2.4/lib/datanucleus-core-3.2.10.jar."
16/02/16 17:00:06 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is
already registered. Ensure you dont have multiple JAR versions of the same plug
in in the classpath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1
-bin-hadoop2.4/lib/datanucleus-rdbms-3.2.9.jar" is already registered, and you a
re trying to register an identical plugin located at URL "file:/C:/Users/Adminis
trator/Downloads/spark-1.5.1-bin-hadoop2.4/bin/./lib/datanucleus-rdbms-3.2.9.ja
r."
16/02/16 17:00:06 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is alr
eady registered. Ensure you dont have multiple JAR versions of the same plugin i
n the classpath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1-bin
-hadoop2.4/lib/datanucleus-api-jdo-3.2.6.jar" is already registered, and you are
trying to register an identical plugin located at URL "file:/C:/Users/Administr
ator/Downloads/spark-1.5.1-bin-hadoop2.4/bin/./lib/datanucleus-api-jdo-3.2.6.ja
r."
16/02/16 17:00:07 WARN Connection: BoneCP specified but not present in CLASSPATH
(or one of dependencies)
16/02/16 17:00:07 WARN Connection: BoneCP specified but not present in CLASSPATH
(or one of dependencies)
16/02/16 17:00:15 WARN ObjectStore: Version information not found in metastore.
hive.metastore.schema.validation is not enabled so recording the schema versio
n 1.2.0
16/02/16 17:00:15 WARN ObjectStore: Failed to get database default, returning No
SuchObjectException
16/02/16 17:00:16 WARN : Your hostname, lenovo-PC resolves to a loopback/non-rea
chable address: fe80:0:0:0:0:5efe:c0a8:78%22, but we couldn't find any external
IP address!
16/02/16 17:00:19 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/02/16 17:00:19 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
```

If you will get any exception (e.g.: `NullPointerException`) while executing spark-shell, please follow the following steps:

The exception is often caused by a missing `winutils.exe` file that Spark needs in order to initialize the Hive context, which in turn depends upon Hadoop, which requires native libraries on Windows to work properly. Unfortunately, this happens even if you are using Spark in local mode without utilizing any of the HDFS features directly.

To resolve this problem, you need to:

Download the 64-bit winutils.exe

- <https://github.com/steveloughran/winutils/tree/master/hadoop-2.6.0/bin>

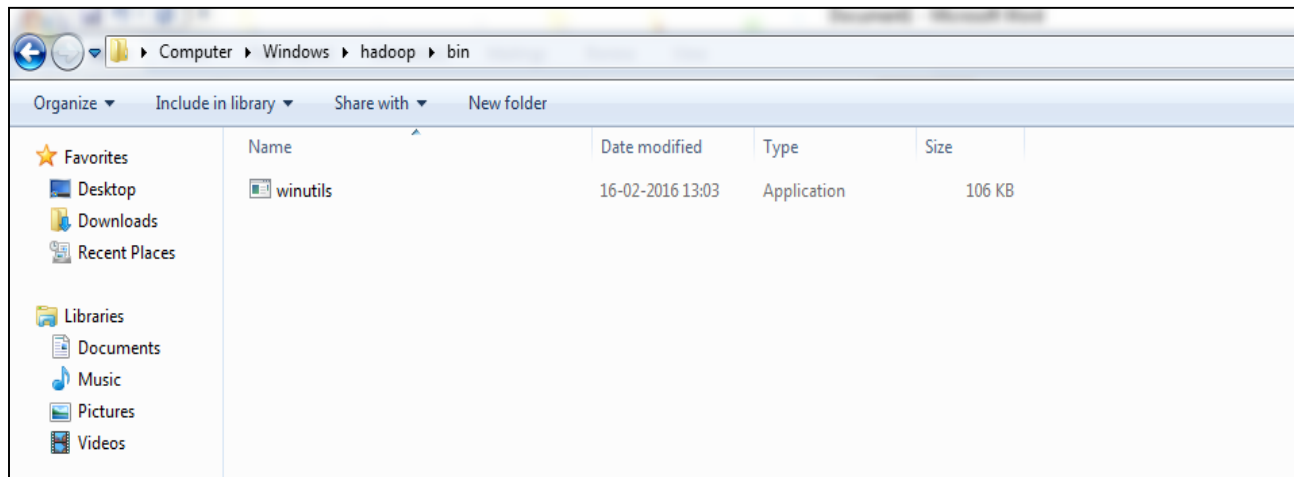
Direct download link:

- <https://github.com/steveloughran/winutils/raw/master/hadoop-2.6.0/bin/winutils.exe>

Note:

There is a different winutils.exe file for the 32-bit Windows version and it will work on the 64-bit Windows Version

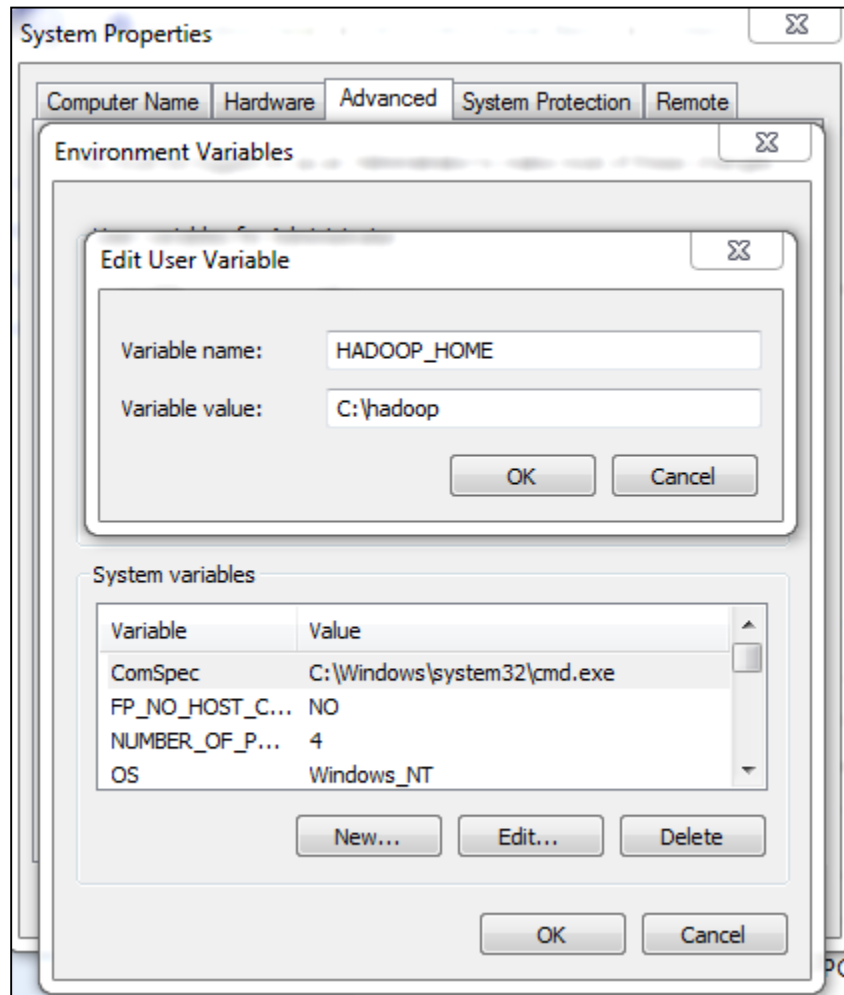
- Copy the downloaded file winutils.exe into a folder like `c:\hadoop\bin` (or `c:\spark\hadoop\bin`)



- Set the environment variable HADOOP\_HOME to point to the above directory but without \bin.

For Example:

If you copied the winutils.exe to C:\hadoop\bin, set `HADOOP_HOME=C:\hadoop`



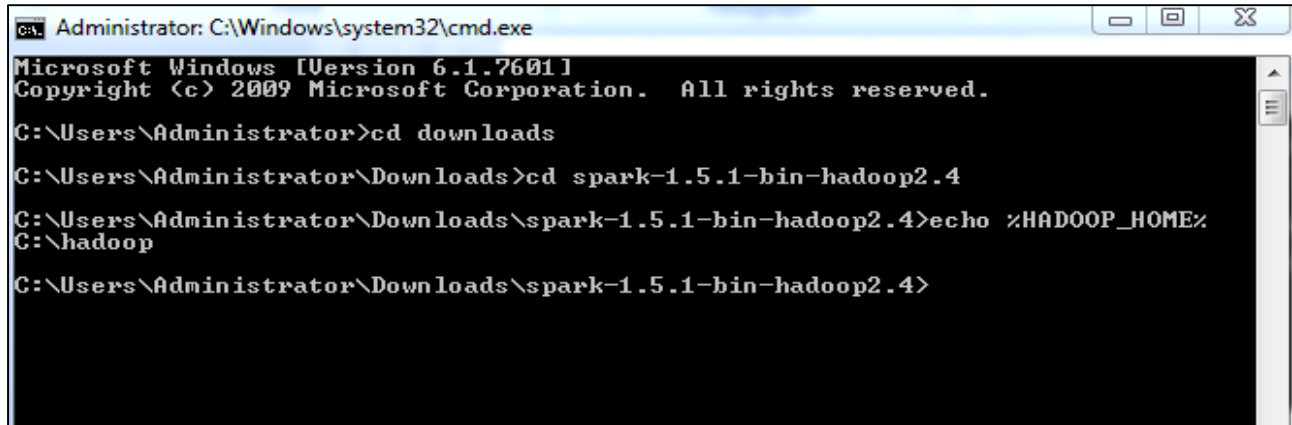
- In Command Prompt change the directory:

```
Administrator: C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd Downloads
C:\Users\Administrator\Downloads>cd spark-1.5.1-bin-hadoop2.4
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>
```

- Check that the environment variable HADOOP\_HOME is set properly by opening Command Prompt and running the following

```
echo %HADOOP_HOME%
```

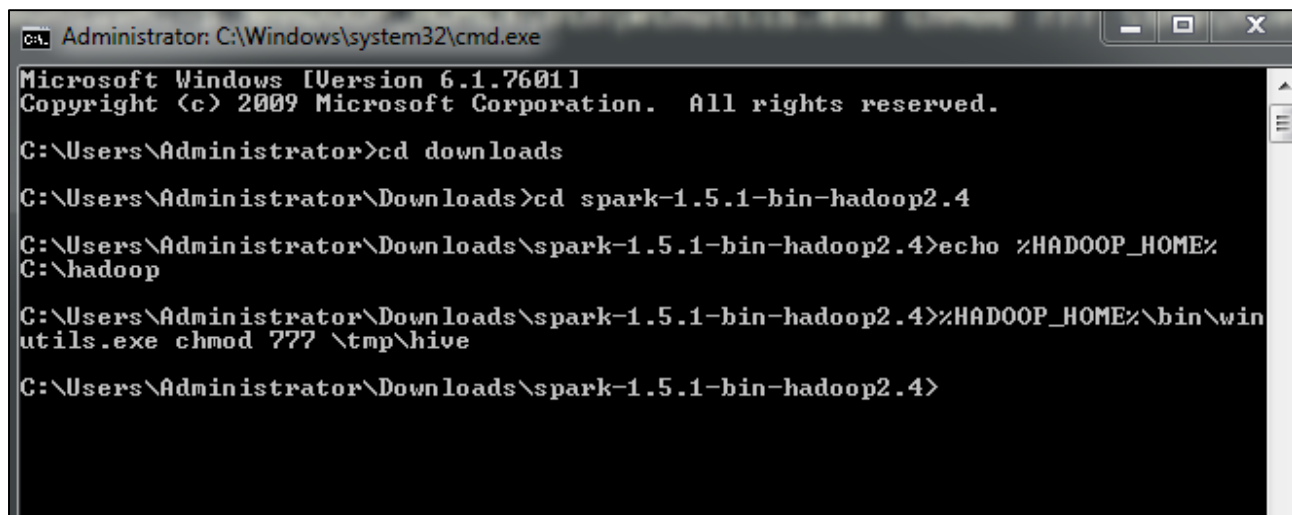


```
Administrator: C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd downloads
C:\Users\Administrator\Downloads>cd spark-1.5.1-bin-hadoop2.4
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>echo %HADOOP_HOME%
C:\hadoop
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>
```

- Set permissions:

```
%HADOOP_HOME%\bin\winutils.exe chmod 777 \tmp\hive
```



```
Administrator: C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd downloads
C:\Users\Administrator\Downloads>cd spark-1.5.1-bin-hadoop2.4
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>echo %HADOOP_HOME%
C:\hadoop
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>%HADOOP_HOME%\bin\win
utils.exe chmod 777 \tmp\hive
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4>
```

- bin\spark-shell

```
C:\Users\Administrator> C:\Windows\system32\cmd.exe - bin\spark-shell
C:\Users\Administrator\Downloads\spark-1.5.1-bin-hadoop2.4> bin\spark-shell
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.lib.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.properties
To adjust logging level use sc.setLogLevel("INFO")
Welcome to

  _ _ _ _ _
 _ _ _ _ _ version 1.5.1
  _ _ _ _ _

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_79)
Type in expressions to have them evaluated.
Type :help for more information.
16/02/16 17:43:27 WARN MetricsSystem: Using default name DAGScheduler for source
because spark.app.id is not set.
Spark context available as sc.
16/02/16 17:43:29 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
istered. Ensure you dont have multiple JAR versions of the same plugin in the cl
asspath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1-bin-hadoop2
.4/bin/./lib/datanucleus-core-3.2.10.jar" is already registered, and you are tr
ying to register an identical plugin located at URL "file:/C:/Users/Administrato
r/Downloads/spark-1.5.1-bin-hadoop2.4/lib/datanucleus-core-3.2.10.jar."
16/02/16 17:43:29 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is
already registered. Ensure you dont have multiple JAR versions of the same plug
in in the classpath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1
-bin-hadoop2.4/lib/datanucleus-rdbms-3.2.9.jar" is already registered, and you a
re trying to register an identical plugin located at URL "file:/C:/Users/Adminis
trator/Downloads/spark-1.5.1-bin-hadoop2.4/bin/./lib/datanucleus-rdbms-3.2.9.ja
r."
16/02/16 17:43:30 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is alr
eady registered. Ensure you dont have multiple JAR versions of the same plugin i
n the classpath. The URL "file:/C:/Users/Administrator/Downloads/spark-1.5.1-bin
-hadoop2.4/lib/datanucleus-api-jdo-3.2.6.jar" is already registered, and you are
trying to register an identical plugin located at URL "file:/C:/Users/Administr
ator/Downloads/spark-1.5.1-bin-hadoop2.4/bin/./lib/datanucleus-api-jdo-3.2.6.ja
r."
16/02/16 17:43:30 WARN Connection: BoneCP specified but not present in CLASSPATH
(or one of dependencies)
16/02/16 17:43:30 WARN Connection: BoneCP specified but not present in CLASSPATH
(or one of dependencies)
16/02/16 17:43:37 WARN ObjectStore: Version information not found in metastore.
hive.metastore.schema.verification is not enabled so recording the schema versio
n 1.2.0
16/02/16 17:43:37 WARN ObjectStore: Failed to get database default, returning No
SuchObjectException
16/02/16 17:43:38 WARN : Your hostname, lenovo-PC resolves to a loopback/non-rea
chable address: fe80:0:0:0:0:5efe:c0a8:72%22, but we couldn't find any external
IP address!
16/02/16 17:43:41 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
16/02/16 17:43:42 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
```