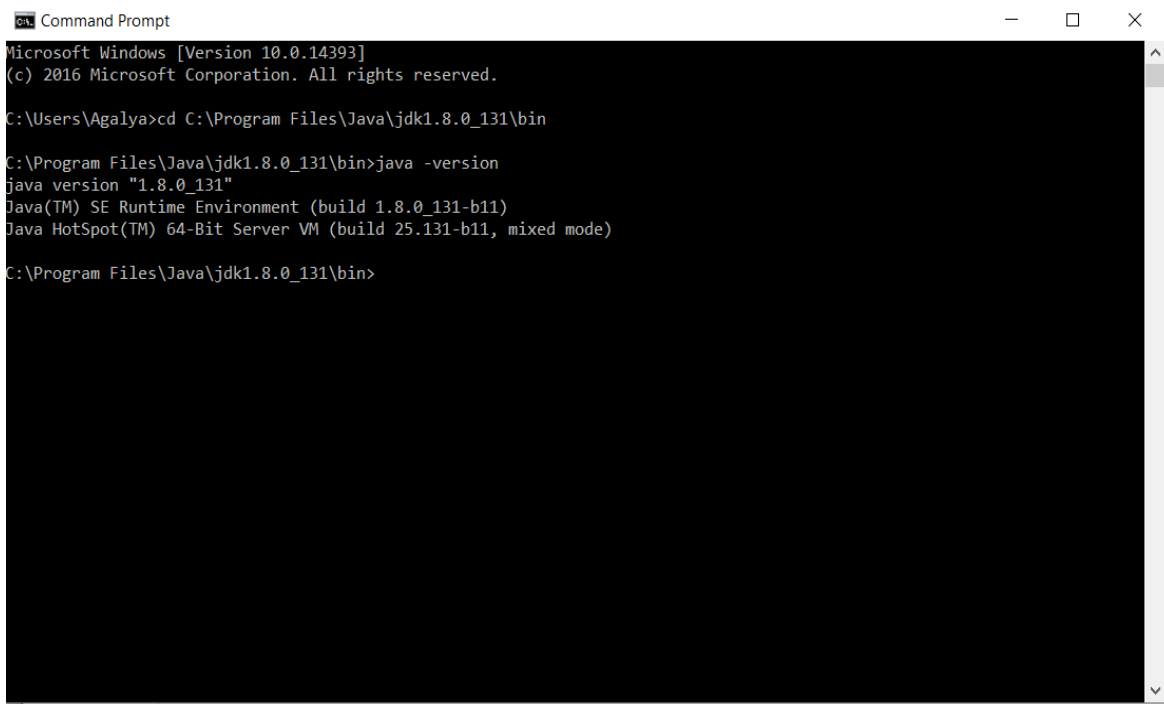


INSTALLING SPARK IN WINDOWS

1. **Install Scala:** Download Scala from the link: <https://www.scala-lang.org/download/2.11.8.html>
 - a. Set environmental variables:
 - i. User variable:
 - Variable: SCALA_HOME;
 - Value: C:\Program Files\scala-2.12.2\scala
 - ii. System variable:
 - Variable: PATH
 - Value: C:\Program Files\scala-2.12.2\scala\bin
 - b. Check in on command prompt

2. **Install Java 8:** Download Java from the link: <https://java.com/en/download>
 - a. Set environmental variables:
 - i. User variable:
 - Variable: JAVA_HOME

- Value: C:\Program Files\Java\jdk1.8.0_91
- ii. System variable:
 - Variable: PATH
 - Value: C:\Program Files\Java\jdk1.8.0_91\bin
- b. Check the cmd:



```
Command Prompt
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\Agalya>cd C:\Program Files\Java\jdk1.8.0_131\bin

C:\Program Files\Java\jdk1.8.0_131\bin>java -version
java version "1.8.0_131"
Java(TM) SE Runtime Environment (build 1.8.0_131-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.131-b11, mixed mode)

C:\Program Files\Java\jdk1.8.0_131\bin>
```

3. Install Eclipse Mars. Download it from the link: <https://eclipse.org/downloads> and extract it into C drive.

- a. Set environmental variables:
 - i. User variable:
 - Variable: ECLIPSE_HOME
 - Value: C:\eclipse
 - ii. System variable:
 - Variable: PATH
 - Value: C:\eclipse\bin

4. Install Spark 1.6.1. Download it from the following link: <http://spark.apache.org/downloads.html> and extract it into C drive, such as C:\spark.

Download Apache Spark™

1. Choose a Spark release: **2.1.0 (Dec 28 2016)**
2. Choose a package type: **Pre-built for Hadoop 2.7 and later**
3. Choose a download type: **Direct Download**
4. Download Spark: **spark-2.1.0-bin-hadoop2.7.tgz**
5. Verify this release using the **2.1.0 signatures and checksums** and **project release KEYS**.

Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

Link with Spark

Spark artifacts are hosted in [Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.11
version: 2.1.0
```

Latest News

- Spark Summit (June 5-7th, 2017, San Francisco) agenda posted (Mar 31, 2017)
- Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)
- Spark 2.1.0 released (Dec 28, 2016)
- Spark wins CloudSort Benchmark as the most efficient engine (Nov 15, 2016)

[Archive](#)

Download Spark

Built-in Libraries:

- [SQL and DataFrames](#)
- [Spark Streaming](#)

a. Set environmental variables:

i. User variable:

- Variable: SPARK_HOME
- Value: C:\spark-2.1.0-bin-hadoop2.7

ii. System variable:

- Variable: C:\spark-2.1.0-bin-hadoop2.7\bin

5.Download Windows Utilities: Download it from the link:

<https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe> And paste it in C:\spark\spark-1.7.1-bin-hadoop2.7\bin

6.Got an error when execute in the command prompt as below:

```
Command Prompt
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\Agalya>cd c:\

c:\>cd spark

c:\spark>spark-shell
'cmd' is not recognized as an internal or external command,
operable program or batch file.

c:\spark>pyspark
'cmd' is not recognized as an internal or external command,
operable program or batch file.

c:\spark>
```

7. Set path in environmental variable as

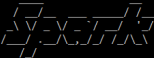
I. System Variable:

PATH: "C:\Windows\system32"

And check in command propmt :

```
Command Prompt - spark-shell
at org.apache.spark.sql.sparkSession.sharedState$lzycompute(SparkSession.scala:101)
at org.apache.spark.sql.sparkSession.sharedState(SparkSession.scala:100)
at org.apache.spark.sql.internal.SessionState.<init>(SessionState.scala:157)
at org.apache.spark.sql.hive.HiveSessionState.<init>(HiveSessionState.scala:32)
... 63 more
Caused by: java.lang.reflect.InvocationTargetException: java.lang.reflect.InvocationTargetException: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be wr
ble. Current permissions are: -----
at sun.reflect.NativeConstructorAccessorImpl.newInstance(Native Method)
at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
at org.apache.spark.sql.internal.SharedState$.org$spark$sql$internal$SharedState$$reflect(SharedState.scala:166)
... 71 more
Caused by: java.lang.reflect.InvocationTargetException: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: -----
at sun.reflect.NativeConstructorAccessorImpl.newInstance(Native Method)
at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
at org.apache.spark.sql.hive.client.IsolatedClientLoader.createClient(IsolatedClientLoader.scala:264)
at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:366)
at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:270)
at org.apache.spark.sql.hive.HiveExternalCatalog.<init>(HiveExternalCatalog.scala:65)
... 76 more
Caused by: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: -----
at org.apache.hadoop.hive.q1.session.SessionState.start(SessionState.java:522)
at org.apache.spark.sql.hive.client.HiveClientImpl.<init>(HiveClientImpl.scala:192)
... 84 more
Caused by: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: -----
at org.apache.hadoop.hive.q1.session.SessionState.createRootHDFSDir(SessionState.java:612)
at org.apache.hadoop.hive.q1.session.SessionState.createSessionDirs(SessionState.java:554)
at org.apache.hadoop.hive.q1.session.SessionState.start(SessionState.java:508)
... 85 more
<console>:14: error: not found: value spark
import spark.implicitly._
^
<console>:14: error: not found: value spark
import spark.sql
^

Welcome to

 version 2.1.0

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_131)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

8. Execute Spark on cmd, see below:

```
Command Prompt - spark-shell

... 76 more
Caused by: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: -----
    at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:522)
    at org.apache.spark.sql.hive.client.HiveClientImpl.<init>(HiveClientImpl.scala:192)
    ... 84 more
Caused by: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: -----
    at org.apache.hadoop.hive.ql.session.SessionState.createRootHDFSDir(SessionState.java:612)
    at org.apache.hadoop.hive.ql.session.SessionState.createSessionDirs(SessionState.java:554)
    at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:508)
    ... 85 more
<console>:14: error: not found: value spark
    import spark.implicitly._
           ^
<console>:14: error: not found: value spark
    import spark.sql
           ^
Welcome to
  ____
 /  __ \
/   /  \
/_____/    version 2.1.0

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_131)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

9. Install Maven 3.3. Download Apache-Maven-3.3.9 from the link:

<https://maven.apache.org/download.cgi> And extract it into C drive, such as C:\apache-maven-3.3.9

a. Set Environmental variables:

i. User variable

- Variable: MAVEN_HOME
- Value: D:\apache-maven-3.3.9

ii. System variable

- Variable: Path
- Value: C:\apache-maven-3.3.9\bin

b. Check on cmd, see below

```
Command Prompt
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\Agalya>cd C:\apache-maven-3.3.9\bin

C:\apache-maven-3.3.9\bin>mvn
[INFO] Scanning for projects...
[INFO] -----
[INFO] BUILD FAILURE
[INFO] -----
[INFO] Total time: 0.062 s
[INFO] Finished at: 2017-05-01T11:00:24+05:30
[INFO] Final Memory: 5M/123M
[INFO] -----
[ERROR] No goals have been specified for this build. You must specify a valid lifecycle phase or a goal in the format <plugin-prefix>:<goal> or <plugin-group-id>:<plugin-artifact-id>[:<plugin-version>]:<goal>. Available lifecycle phases are : validate, initialize, generate-sources, process-sources, generate-resources, process-resources, compile, process-classes, generate-test-sources, process-test-sources, generate-test-resources, process-test-resources, test-compile, process-test-classes, test, prepare-package, package, pre-integration-test, integration-test, post-integration-test, verify, install, deploy, pre-clean, clean, post-clean, pre-site, site, post-site, site-deploy. -> [Help 1]
[ERROR]
[ERROR] To see the full stack trace of the errors, re-run Maven with the -e switch.
[ERROR] Re-run Maven using the -X switch to enable full debug logging.
[ERROR]
[ERROR] For more information about the errors and possible solutions, please read the following articles:
[ERROR] [Help 1] http://cwiki.apache.org/confluence/display/MAVEN/NoGoalSpecifiedException

C:\apache-maven-3.3.9\bin>
```

10.Connect sparkR in command prompt.

- set the path in command prompt: “C:\Program Files\R\R-3.3.1\bin\x64” and run “sparkR” in cmd:

Rterm (64-bit)

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Spark package found in SPARK_HOME: C:\spark-2.1.0-bin-hadoop2.6\bin\..

Launching java with spark-submit command C:\spark-2.1.0-bin-hadoop2.6\bin\..\bin\spark-submit2.cmd "sparkr-shell" C:\Users\Agalya\AppData\Local\Temp\RtmpA1nGs6\backend_port3fc47cc82238b

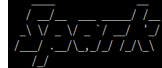
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

17/05/03 09:48:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Welcome to

 version 2.1.0

SparkSession available as 'spark'.

> x=7

> x

[1] 7

> .