

Installing Spark Pre-built for Hadoop version 2.6 and later

You still need to install Ubuntu on Virtual Machine and Java using same commands provided earlier by Ashwin(Step 1-5 remains same).

1. Download VirtualBox from Oracle Website

<https://www.virtualbox.org/wiki/Downloads>

2. Install Oracle VirtualBox like any other software.

3. Download Ubuntu Desktop Image from the following URL

<http://www.ubuntu.com/download>

4. Use the instructions in the below link to setup Ubuntu in VirtuaBox

<http://www.wikihow.com/Install-Ubuntu-on-VirtualBox>

5. After Installation try to installing Ubuntu, open Terminal and execute the following operations

a. Login as root user using the command – “su root”

b. # sudo apt-get update

c. Install Java in Ubuntu

```
# sudo add-apt-repository ppa:webupd8team/java
```

```
# sudo apt-get update
```

```
# sudo apt-get install oracle-java7-installer
```

d. Set Java environment variable

```
# vi /etc/environment
```

Paste the following line in that document

```
export JAVA_HOME=/usr/lib/jvm/java-7-oracle
```

Press ESC key and type the following command to save the file in VI Editor

```
:wq
```

6. Download the latest Spark version 1.3.0 from <http://spark.apache.org/downloads.html>

a. Select the Pre-built for Hadoop 2.6 and later

The screenshot shows the Apache Spark download page in a web browser. The page title is "Download Spark". It states "The latest release of Spark is Spark 1.3.1, released on April 17, 2015 (release notes) (git tag)". There are five steps to follow:

1. Chose a Spark release: 1.3.1 (Apr 17 2015)
2. Chose a package type: Pre-built for Hadoop 2.6 and later
3. Chose a download type: Pre-built for Hadoop 2.6 and later
4. Download Spark: spark
5. Verify this release using: scala 2.11 support.

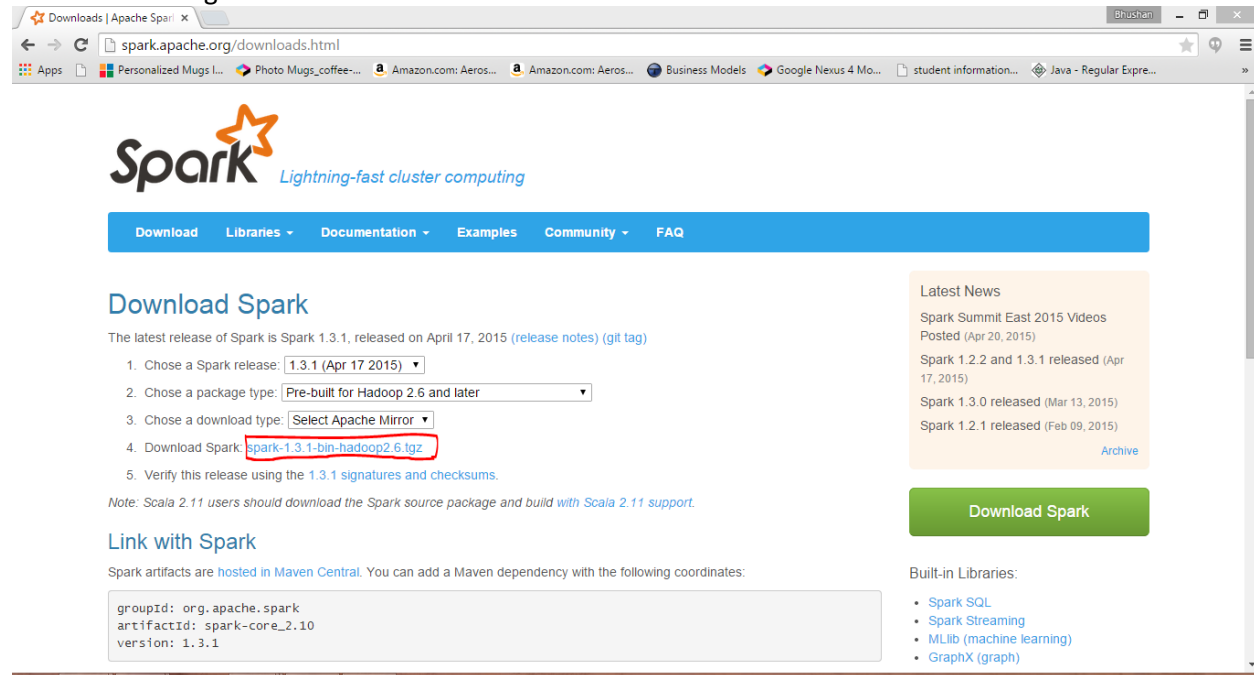
Below the steps, there is a note: "Note: Scala 2.11 users should use the Pre-built for CDH 4".

Under "Link with Spark", it says "Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:"

```
groupId: org.apache.spark
artifactId: spark-core_2.10
version: 1.3.1
```

On the right side, there is a "Latest News" section with links to "Spark Summit East 2015 Videos", "Spark 1.2.2 and 1.3.1 released", "Spark 1.3.0 released", and "Spark 1.2.1 released". There is also a "Download Spark" button and a "Built-in Libraries" section listing "Spark SQL", "Spark Streaming", "MLlib (machine learning)", and "GraphX (graph)".

Download the .tgz file.



The screenshot shows the Apache Spark download page in a web browser. The page title is "Download Spark". It lists the latest release as Spark 1.3.1, released on April 17, 2015. The steps to download are: 1. Choose a Spark release: 1.3.1 (Apr 17 2015). 2. Choose a package type: Pre-built for Hadoop 2.6 and later. 3. Choose a download type: Select Apache Mirror. 4. Download Spark: spark-1.3.1-bin-hadoop2.6.tgz (highlighted with a red box). 5. Verify this release using the 1.3.1 signatures and checksums. A note mentions Scala 2.11 users should download the Spark source package and build with Scala 2.11 support. On the right, there is a "Latest News" section with links to Spark Summit East 2015 Videos, Spark 1.2.2 and 1.3.1 released, Spark 1.3.0 released, and Spark 1.2.1 released. A green "Download Spark" button is present. Below the download steps, there is a "Link with Spark" section showing Maven coordinates: groupId: org.apache.spark, artifactId: spark-core_2.10, version: 1.3.1. A "Built-in Libraries" section lists Spark SQL, Spark Streaming, MLlib (machine learning), and GraphX (graph).

Download Spark

Latest News

- Spark Summit East 2015 Videos Posted (Apr 20, 2015)
- Spark 1.2.2 and 1.3.1 released (Apr 17, 2015)
- Spark 1.3.0 released (Mar 13, 2015)
- Spark 1.2.1 released (Feb 09, 2015)

Download Spark

Built-in Libraries:

- Spark SQL
- Spark Streaming
- MLlib (machine learning)
- GraphX (graph)

b. extract the file using the following command

```
# tar zvfz spark-1.3.0.tgz
```

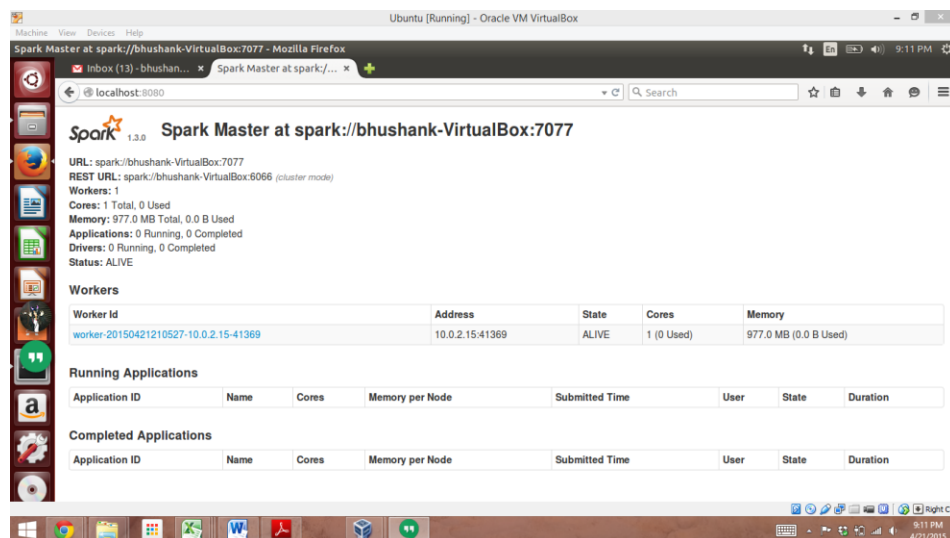
c. get inside the extracted folder using the following command

```
# cd spark-1.3.0-bin-hadoop2.6
```

d. Goto "sbin" folder and execute the following commands and check for **localhost:8080**

```
# ./stop-all.sh
```

```
# ./start-all.sh
```



The screenshot shows the Spark Master web interface in a web browser. The URL is localhost:8080. The page displays the Spark Master status for spark://bhushank-VirtualBox:7077. It shows the REST URL, the number of workers (1), and the status (ALIVE). A table lists the workers with columns: Worker ID, Address, State, Cores, and Memory. The table shows one worker with ID worker-20150421210527-10.0.2.15-41369, Address 10.0.2.15:41369, State ALIVE, 1 (0 Used) Cores, and 977.0 MB (0.0 B Used) Memory. Below the workers table, there are sections for Running Applications and Completed Applications, each with a table showing Application ID, Name, Cores, Memory per Node, Submitted Time, User, State, and Duration.

Spark Master at spark://bhushank-VirtualBox:7077

URL: spark://bhushank-VirtualBox:7077
REST URL: spark://bhushank-VirtualBox:6066 (cluster mode)

Workers: 1
Cores: 1 Total, 0 Used
Memory: 977.0 MB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Worker ID	Address	State	Cores	Memory
worker-20150421210527-10.0.2.15-41369	10.0.2.15:41369	ALIVE	1 (0 Used)	977.0 MB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

d. Get back to spark-1.3.0-bin-hadoop2.6 folder and execute following command
./bin/pyspark

7. To remove logging in pyspark, execute the following command.

a. Get into "conf" folder

cp log4j.properties.template log4j.properties

vi log4j.properties

Change

log4j.rootCategory = INFO, console

to

log4j.rootCategory = ERROR, console

Happy Coding !!