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

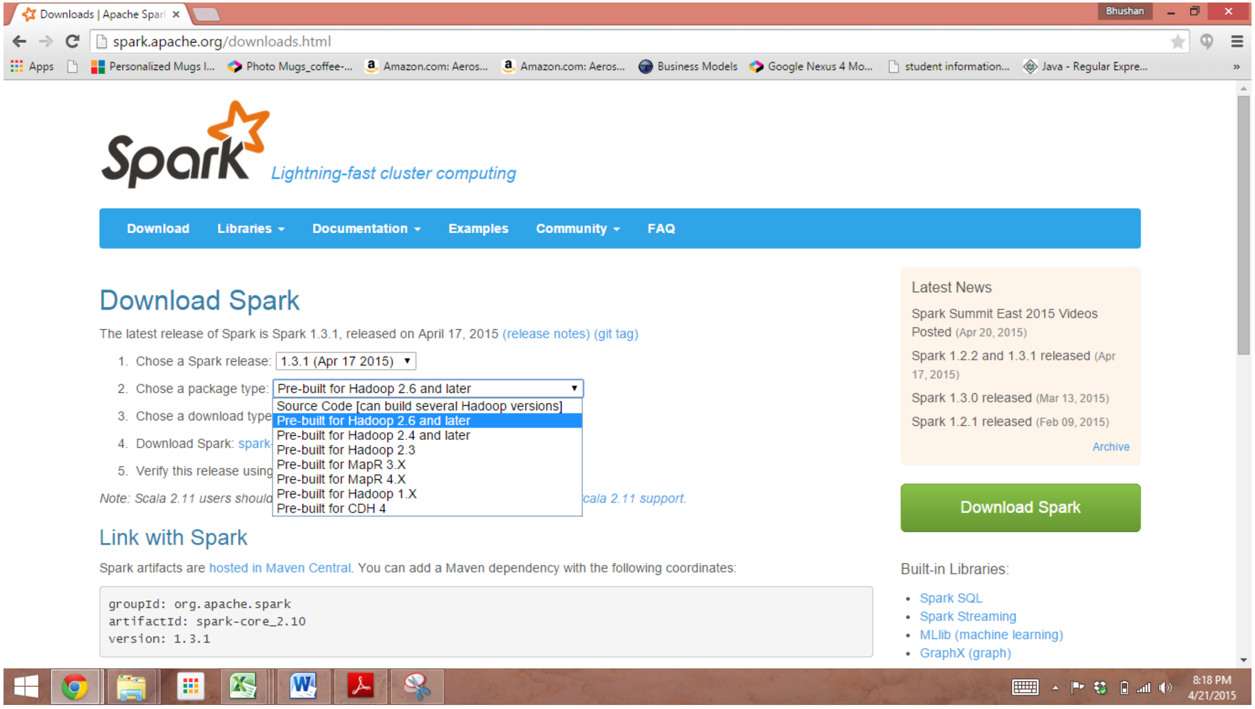
```
export JAVA_HOME=/usr/lib/jvm/java-7-oracle
```

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



Download Spark

The latest release of Spark is Spark 1.3.1, released on April 17, 2015 (release notes) (git tag)

1. Chose a Spark release: 1.3.1 (Apr 17 2015)
2. Chose a package type: Pre-built for Hadoop 2.6 and later
   - Source Code [can build several Hadoop versions]
   - Pre-built for Hadoop 2.6 and later
   - Pre-built for Hadoop 2.4 and later
   - Pre-built for Hadoop 2.3
   - Pre-built for MapR 3.X
   - Pre-built for MapR 4.X
   - Pre-built for Hadoop 1.X
   - Pre-built for CDH 4
3. Chose a download type
4. Download Spark: spark
5. Verify this release using

Note: Scala 2.11 users should ... cala 2.11 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_2.10
version: 1.3.1
```

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)

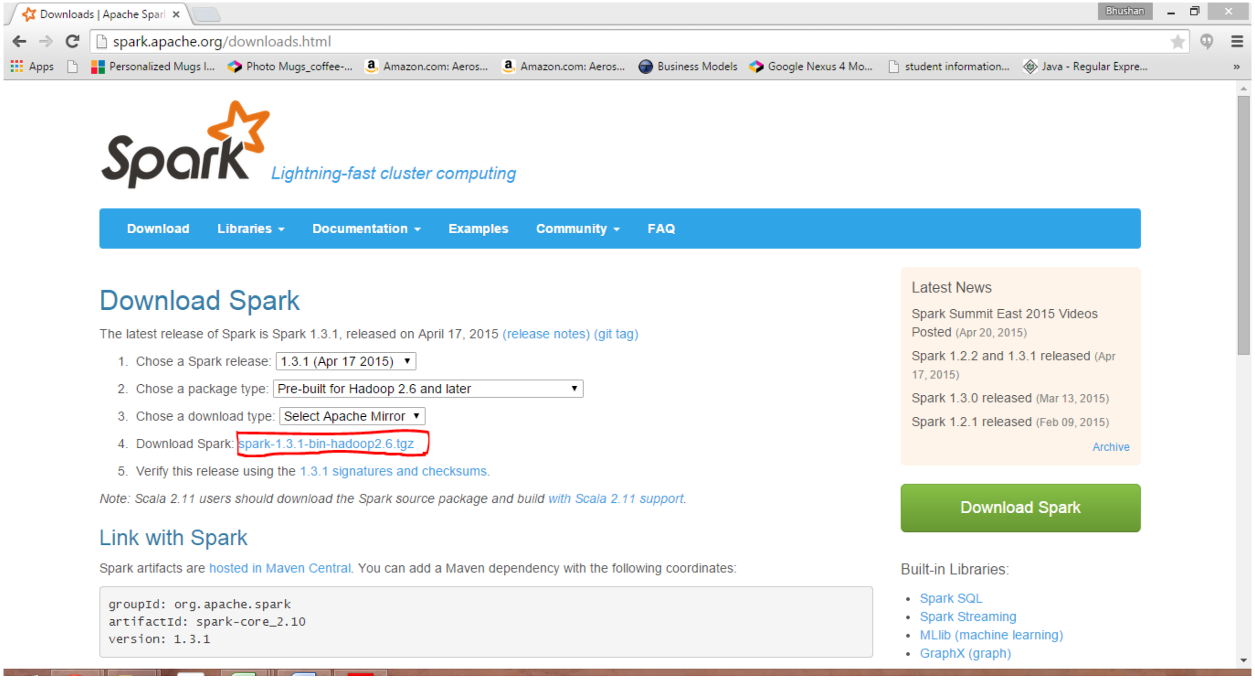
Spark 1.2.1 released (Feb 09, 2015)

Archive

Built-in Libraries:

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

Download the .tgz file.



Download Spark

The latest release of Spark is Spark 1.3.1, released on April 17, 2015 (release notes) (git tag)

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: Select Apache Mirror
4. Download Spark: spark-1.3.1-bin-hadoop2.6.tgz
5. Verify this release using the 1.3.1 signatures and checksums.

Note: Scala 2.11 users should download the Spark source package and build with Scala 2.11 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_2.10
version: 1.3.1
```

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)

Archive

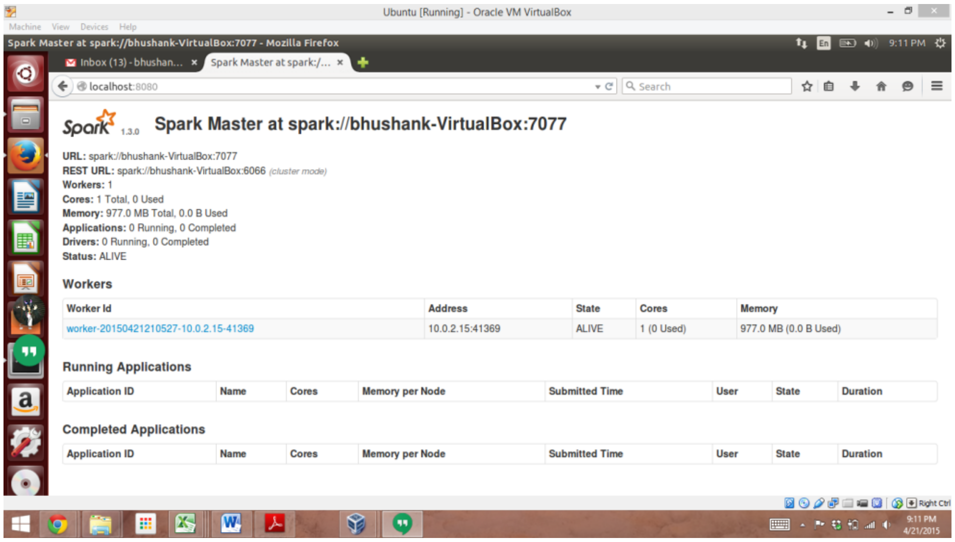
Built-in Libraries:

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

b. extract the file using the following command
# tar zvfx 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


Ubuntu [Running] - Oracle VM VirtualBox
Machine
View
Devices
Help
Spark Master at spark://bhushank-VirtualBox:7077 - Mozilla Firefox
9:11 PM
Inbox (13) - bhushan...
Spark Master at spark:/...
localhost:8080
Search
Spark 1.3.0

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

Workers

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 !!