

# Spark Installation Guide

Ladle Patel



# Java Installation

```
ubuntu@ip-172-31-56-161:~$ sudo add-apt-repository ppa:webupd8team/java -y
ubuntu@ip-172-31-56-161:~$ sudo apt-get update
ubuntu@ip-172-31-56-161:~$ sudo apt-get install oracle-java8-installer
```

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

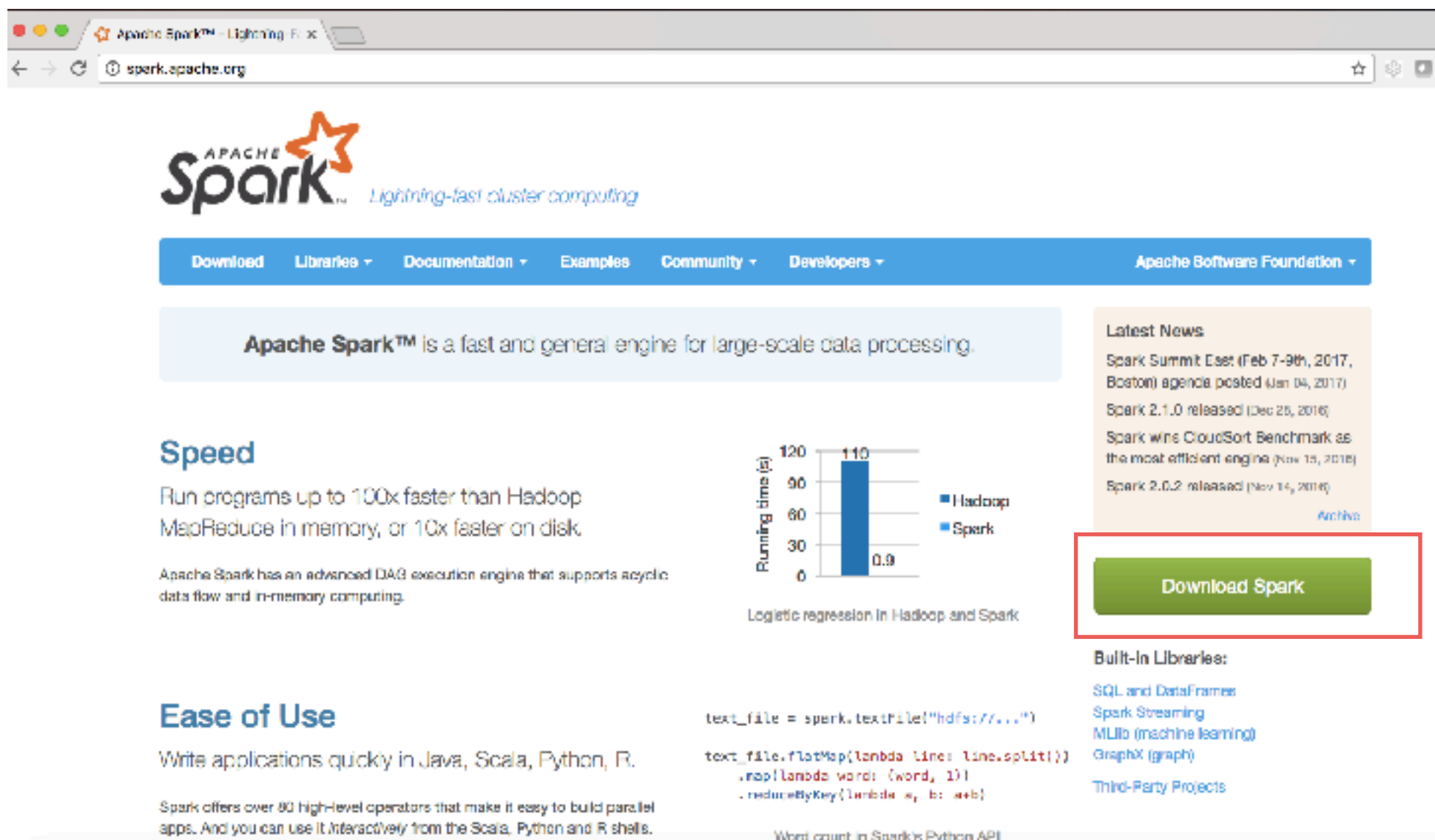
`sudo apt-get update`

`sudo apt-get install oracle-java8-installer`

`java -version`

# Spark Installation

- Goto <http://spark.apache.org/>
- Click on Downloads (Marked in below image)

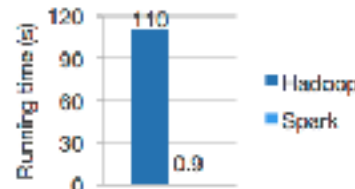


Apache Spark™ is a fast and general engine for large-scale data processing.

## Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

## Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala, Python and R shells.

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

## Latest News

- Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)
- Spark 2.1.0 released (Dec 25, 2016)
- Spark wins CloudSort Benchmark as the most efficient engine (Nov 15, 2016)
- Spark 2.0.2 released (Nov 14, 2016)

[Archive](#)

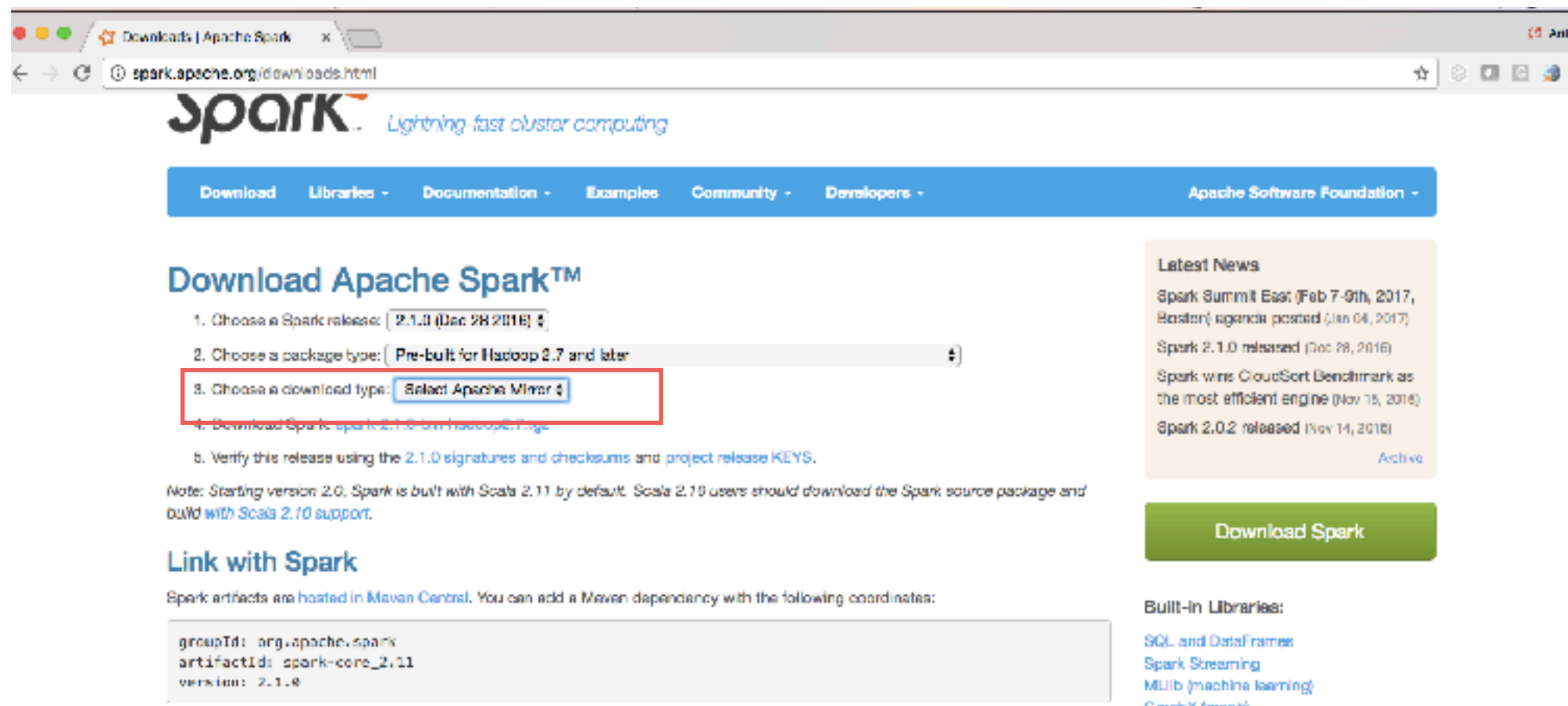
## Built-In Libraries:

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

## Third-Party Projects

# Spark Installation Cont..

- Choose a download type as “Select Apache mirror”(Chose Direct Download option if you are installing in your laptop or server which has GUI).



The screenshot shows the Apache Spark download page in a web browser. The browser's address bar displays `spark.apache.org/downloads.html`. The page features the Spark logo and the tagline "Lightning fast cluster computing". A navigation bar includes links for Download, Libraries, Documentation, Examples, Community, Developers, and Apache Software Foundation. The main content area is titled "Download Apache Spark™" and contains a list of steps for downloading Spark. Step 3, "Choose a download type:", is highlighted with a red rectangle, showing the "Select Apache Mirror" option selected. To the right, a "Latest News" section lists recent events and releases. At the bottom right, there is a green "Download Spark" button and a section for "Built-in Libraries".

**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: `Select Apache Mirror`
4. Download Spark `spark-2.1.0-bin-hadoop2.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 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)
- Spark 2.0.2 released (Nov 14, 2016)

[Archive](#)

**Download Spark**

**Built-in Libraries:**

- [SQL and DataFrames](#)
- [Spark Streaming](#)
- [MLlib \(machine learning\)](#)
- [GraphX \(graphs\)](#)

# Spark Installation Cont..

- Copy below suggested URL and paste using by preceding wget command at terminal.

A screenshot of a web browser showing the Apache Spark download page. The browser's address bar displays the URL: www.apache.org/dyn/closer.lua/spark/spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz. The page features the Apache Software Foundation logo on the left and a navigation menu on the right with links like "Home", "About", "Projects", "People", "Get Involved", "Download", and "Support Apache". Below the logo, a red-bordered box highlights the text: "We suggest the following mirror site for your download:". Below this box, the URL http://mirror.fibergrid.in/apache/spark/spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz is displayed. Further down, a section titled "HTTP" shows the same URL. A disclaimer at the bottom states: "Other mirror sites are suggested below. Please use the backup mirrors only to download PGP and MD5 signatures to verify your downloads or if no other mirrors are working."

# Spark Installation Cont..

```
ubuntu@ip-172-31-56-161:~$ wget http://mirror.fibergrid.in/apache/spark/spark-2.1.0/  
spark-2.1.0-bin-hadoop2.7.tgz
```

- Download spark tar file

```
wget http://mirror.fibergrid.in/apache/spark/  
spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz
```

- unzip downloaded file

```
tar -xzf spark-2.1.0-bin-hadoop2.7.tgz
```

- Rename file

```
mv spark-2.1.0-bin-hadoop2.7 spark
```

# Set the Path

- Edit bash file.

```
ubuntu@ip-172-31-56-161:~$ vi ~/.bashrc
```

```
# ~/.bashrc: executed by bash(1) for non-login shells.
# See /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
case $- in
  *) ;;
  *) return;;
esac

export SPARK_HOME=/home/ubuntu/spark
export PATH=$PATH:$SPARK_HOME/bin

# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000
HISTFILESIZE=2000

# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize

# If set, the pattern 'we*' used in a pathname expansion context will
```

```
ubuntu@ip-172-31-56-161:~$ vi ~/.bashrc
ubuntu@ip-172-31-56-161:~$ source ~/.bashrc
ubuntu@ip-172-31-56-161:~$
```

# Set the Path Cont..

```
vi ~/.bashrc
```

```
export SPARK_HOME=/home/ubuntu/spark
```

```
export PATH=$PATH:$SPARK_HOME/bin
```

Compile the changes

- Compile changes.

```
source ~/.bashrc
```



# Starting Spark

- Starting Spark with Python

```
ubuntu@ip-172-31-56-161:~$ pyspark █
```

pyspark

- Starting Spark with Scala

```
Last login: Sun Mar 26 12:39:18 2017 from 49.227.66.65  
ubuntu@ip-172-31-56-161:~$ spark-shell █
```

spark-shell



Thank You.