


Assignment 12.3: Spark Installation

Step 1: Go to below URL and select appropriate version of spark and download clicking on the highlighted link.

<https://spark.apache.org/downloads.html>

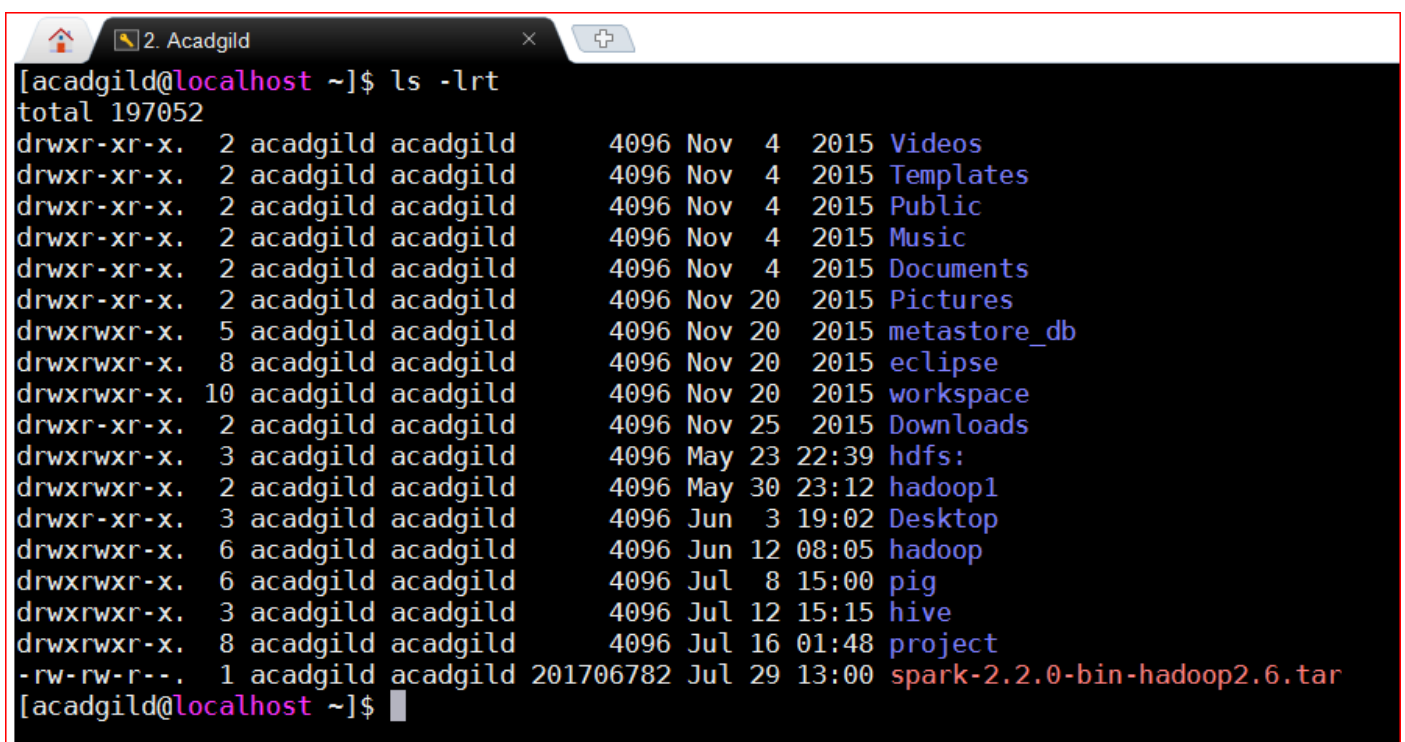


The screenshot shows the Apache Spark download page. At the top is the Apache Spark logo with the tagline "Lightning-fast cluster computing". Below the logo is a navigation bar with links: Download, Libraries, Documentation, Examples, Community, and Developers. The main heading is "Download Apache Spark™". Below this, there are five steps to follow:

1. Choose a Spark release: 2.2.0 (Jul 11 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.6
3. Choose a download type: Direct Download
4. Download Spark: **spark-2.2.0-bin-hadoop2.6.tgz**
5. Verify this release using the 2.2.0 signatures and checksums and project release KEYS.

A note at the bottom states: "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."

Step 2: Move downloaded file to Acadgild sandbox.

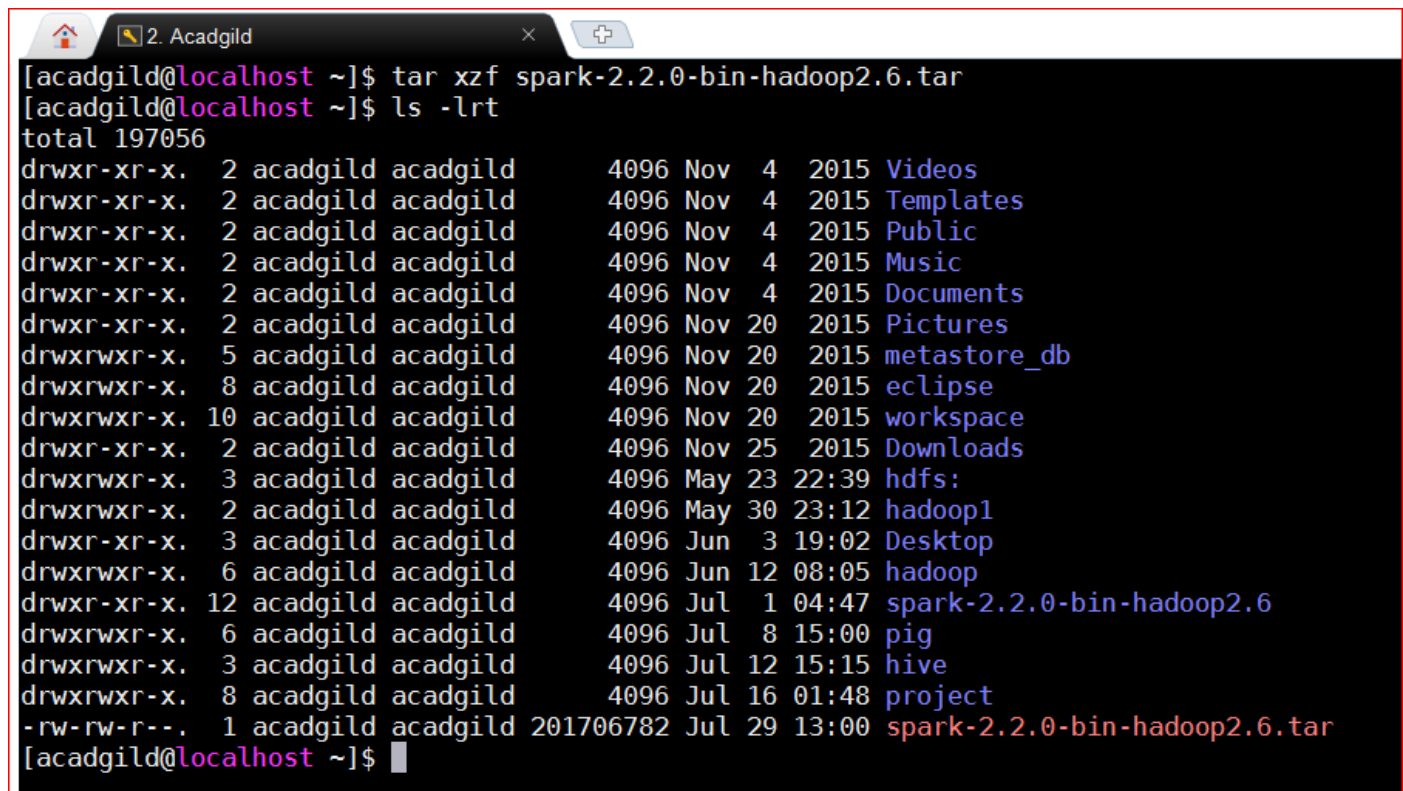


The screenshot shows a terminal window with the command `ls -lrt` executed. The output lists files and directories in the home directory of the user 'acadgild' on 'localhost'. The files are sorted by time, showing various system directories and user files. The file `spark-2.2.0-bin-hadoop2.6.tar` is highlighted in red in the original image, indicating it is the file to be moved to the sandbox.

```
[acadgild@localhost ~]$ ls -lrt
total 197052
drwxr-xr-x. 2 acadgild acadgild 4096 Nov  4  2015 Videos
drwxr-xr-x. 2 acadgild acadgild 4096 Nov  4  2015 Templates
drwxr-xr-x. 2 acadgild acadgild 4096 Nov  4  2015 Public
drwxr-xr-x. 2 acadgild acadgild 4096 Nov  4  2015 Music
drwxr-xr-x. 2 acadgild acadgild 4096 Nov  4  2015 Documents
drwxr-xr-x. 2 acadgild acadgild 4096 Nov 20  2015 Pictures
drwxrwxr-x. 5 acadgild acadgild 4096 Nov 20  2015 metastore_db
drwxrwxr-x. 8 acadgild acadgild 4096 Nov 20  2015 eclipse
drwxrwxr-x. 10 acadgild acadgild 4096 Nov 20  2015 workspace
drwxr-xr-x. 2 acadgild acadgild 4096 Nov 25  2015 Downloads
drwxrwxr-x. 3 acadgild acadgild 4096 May 23 22:39 hdfs:
drwxrwxr-x. 2 acadgild acadgild 4096 May 30 23:12 hadoop1
drwxr-xr-x. 3 acadgild acadgild 4096 Jun  3 19:02 Desktop
drwxrwxr-x. 6 acadgild acadgild 4096 Jun 12 08:05 hadoop
drwxrwxr-x. 6 acadgild acadgild 4096 Jul  8 15:00 pig
drwxrwxr-x. 3 acadgild acadgild 4096 Jul 12 15:15 hive
drwxrwxr-x. 8 acadgild acadgild 4096 Jul 16 01:48 project
-rw-rw-r--. 1 acadgild acadgild 201706782 Jul 29 13:00 spark-2.2.0-bin-hadoop2.6.tar
[acadgild@localhost ~]$
```

Step 3: Unzip/untar the file using below command.

[acadgild@localhost ~]\$ tar xzf spark-2.2.0-bin-hadoop2.6.tar



```
[acadgild@localhost ~]$ tar xzf spark-2.2.0-bin-hadoop2.6.tar
[acadgild@localhost ~]$ ls -lrt
total 197056
drwxr-xr-x.  2 acadgild acadgild   4096 Nov  4  2015 Videos
drwxr-xr-x.  2 acadgild acadgild   4096 Nov  4  2015 Templates
drwxr-xr-x.  2 acadgild acadgild   4096 Nov  4  2015 Public
drwxr-xr-x.  2 acadgild acadgild   4096 Nov  4  2015 Music
drwxr-xr-x.  2 acadgild acadgild   4096 Nov  4  2015 Documents
drwxr-xr-x.  2 acadgild acadgild   4096 Nov 20  2015 Pictures
drwxrwxr-x.  5 acadgild acadgild   4096 Nov 20  2015 metastore_db
drwxrwxr-x.  8 acadgild acadgild   4096 Nov 20  2015 eclipse
drwxrwxr-x. 10 acadgild acadgild   4096 Nov 20  2015 workspace
drwxr-xr-x.  2 acadgild acadgild   4096 Nov 25  2015 Downloads
drwxrwxr-x.  3 acadgild acadgild   4096 May 23 22:39 hdfs:
drwxrwxr-x.  2 acadgild acadgild   4096 May 30 23:12 hadoop1
drwxr-xr-x.  3 acadgild acadgild   4096 Jun  3 19:02 Desktop
drwxrwxr-x.  6 acadgild acadgild   4096 Jun 12 08:05 hadoop
drwxr-xr-x. 12 acadgild acadgild   4096 Jul  1 04:47 spark-2.2.0-bin-hadoop2.6
drwxrwxr-x.  6 acadgild acadgild   4096 Jul  8 15:00 pig
drwxrwxr-x.  3 acadgild acadgild   4096 Jul 12 15:15 hive
drwxrwxr-x.  8 acadgild acadgild   4096 Jul 16 01:48 project
-rw-rw-r--.  1 acadgild acadgild 201706782 Jul 29 13:00 spark-2.2.0-bin-hadoop2.6.tar
[acadgild@localhost ~]$
```

Step 4: Verify the content of directory **spark-2.2.0-bin-hadoop2.6** and try launching spark with python.

```
[acadgild@localhost spark-2.2.0-bin-hadoop2.6]$ ls -lrt
total 104
-rw-r--r-- 1 acadgild acadgild 24645 Jul 1 04:47 NOTICE
drwxr-xr-x 2 acadgild acadgild 4096 Jul 1 04:47 yarn
drwxr-xr-x 2 acadgild acadgild 4096 Jul 1 04:47 sbin
-rw-r--r-- 1 acadgild acadgild 128 Jul 1 04:47 RELEASE
-rw-r--r-- 1 acadgild acadgild 3809 Jul 1 04:47 README.md
drwxr-xr-x 3 acadgild acadgild 4096 Jul 1 04:47 R
drwxr-xr-x 6 acadgild acadgild 4096 Jul 1 04:47 python
drwxr-xr-x 2 acadgild acadgild 4096 Jul 1 04:47 licenses
-rw-r--r-- 1 acadgild acadgild 17881 Jul 1 04:47 LICENSE
drwxr-xr-x 2 acadgild acadgild 12288 Jul 1 04:47 jars
drwxr-xr-x 4 acadgild acadgild 4096 Jul 1 04:47 examples
drwxr-xr-x 5 acadgild acadgild 4096 Jul 1 04:47 data
drwxr-xr-x 2 acadgild acadgild 4096 Jul 1 04:47 conf
drwxr-xr-x 2 acadgild acadgild 4096 Jul 1 04:47 bin
[acadgild@localhost spark-2.2.0-bin-hadoop2.6]$ bin/pyspark
Python 2.6.6 (r266:84292, Jul 23 2015, 15:22:56)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-11)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
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).
17/07/29 13:15:28 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/07/29 13:15:28 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.56.101 instead (on interface eth5)
17/07/29 13:15:28 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
/home/acadgild/spark-2.2.0-bin-hadoop2.6/python/pyspark/context.py:195: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
17/07/29 13:15:37 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.validation is not enabled so recording the schema version 1.2.0
17/07/29 13:15:37 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/07/29 13:15:38 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to

  ____      _
 / ___|  __| | | |
 \___ \  | | | | | |
  ___) | | | | | | |
 |____|_|_|_|_|_|_|_| version 2.2.0

Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkSession available as 'spark'.
>>>
```

Step 5: close the pyspark (cntrl+d) and rename the directory **spark-2.2.0-bin-hadoop2.6** to **spark** for the simplicity.

```
[acadgild@localhost ~]$ mv spark-2.2.0-bin-hadoop2.6 spark
```

Step 6: update **.bash_profile** file with the path of spark bin directory so that pyspark can be launched from any directory. Make sure that you **source .bash_profile** file.

```
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
    . ~/.bashrc
fi

# User specific environment and startup programs

PATH=$PATH:$HOME/bin

SPARK_HOME=/home/acadgild/spark

PATH=$PATH:$SPARK_HOME/bin

export PATH

~
~
~
```

Step 7: Verify the \$PATH.

```
[acadgild@localhost ~]$ source .bash_profile
[acadgild@localhost ~]$ echo $PATH
/usr/local/hive/bin:/usr/local/hive/bin:/usr/lib64/qt-3.3/bin:/usr/local/bin:/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/sbin:/usr/local/java/bin:/usr/local/hadoop-2.6.0/bin:/usr/local/flume/bin:/usr/local/pig/bin:/usr/local/hive/bin:/usr/local/hbase/bin:/usr/local/sqoop/bin:/usr/local/hadoop-2.6.0/sbin:/usr/local/oozie-4.1.0/bin:/home/acadgild/bin:/usr/local/java/bin:/usr/local/hadoop-2.6.0/bin:/usr/local/flume/bin:/usr/local/pig/bin:/usr/local/hive/bin:/usr/local/hbase/bin:/usr/local/sqoop/bin:/usr/local/hadoop-2.6.0/sbin:/usr/local/oozie-4.1.0/bin:/home/acadgild/bin:/home/acadgild/spark/bin
[acadgild@localhost ~]$
```

Step 8: launch spark using pyspark from anywhere.

```
[acagild@localhost ~]$ pyspark
Python 2.6.6 (r266:84292, Jul 23 2015, 15:22:56)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-11)] on linux2
Type "help", "copyright", "credits" or "license()" for more information.
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).
17/07/29 13:28:16 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/07/29 13:28:16 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.56.101 instead (on interface eth5)
17/07/29 13:28:16 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
/home/acagild/spark/python/pyspark/context.py:195: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
17/07/29 13:28:24 ERROR ObjectStore: Version information found in metastore differs 0.14.0 from expected schema version 1.2.0. Schema verification is disabled hive.metastore.schema.verification so setting version.
17/07/29 13:28:24 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to

  ____      _
 / ___|  __| | | |
| |  | |__| | | |
| |  | |__| | | |
| |  | |__| | | |
|_|  |____|_|_|_|

 version 2.2.0

Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkSession available as 'spark'.
>>> █
```

Step 9: launch another acadgild sandbox session and run `jps` command to see spark daemon running.

```
Last login: Sat Jul 29 12:59:39 2017 from 192.168.56.1
[acadgild@localhost ~]$ jps
5107 Jps
2550 SparkSubmit
[acadgild@localhost ~]$
```

Spark Installed successfully.