Installing Hadoop and Spark

Spark uses Hadoop’s client libraries for HDFS and YARN.

Hadoop installation: summary

- Download Hadoop 2.7.2
- Setup passwordless SSH
- Edit config files
- Setup HDFS - format the namenode/create directories for MapReduce
- Start the daemons

Download Hadoop

Download a recent stable release from one of the Apache Download Mirrors.

Unpack the downloaded Hadoop release.

- Pick a target directory for installing the package.
  - Example: I used **/home/emma/hadoop/** as my target.
- Extract the tar.gz file (e.g.*hadoop-2.7.2.tar.gz*) in your target directory
- This will yield a directory structure as shown:

```
emma@horatio:~/hadoop/hadoop-2.7.2> ll
total 56
drwxr-xr-x 2 emma users  4096 Jan 25  2016 bin
drwxr-xr-x 3 emma users  4096 Jan 25  2016 etc
drwxr-xr-x 2 emma users  4096 Jan 25  2016 include
drwxr-xr-x 3 emma users  4096 Jan 25  2016 lib
drwxr-xr-x 2 emma users  4096 Jan 25  2016 libexec
-rw-r--r-- 1 emma users 15429 Jan 25  2016 LICENSE.txt
drwxr-xr-x 3 emma users  4096 Aug 21 14:19 logs
-rw-r--r-- 1 emma users   101 Jan 25  2016 NOTICE.txt
-rw-r--r-- 1 emma users  1366 Jan 25  2016 README.txt
drwxr-xr-x 2 emma users  4096 Jan 25  2016 sbin
drwxr-xr-x 4 emma users  4096 Jan 25  2016 share
emma@horatio:~/hadoop/hadoop-2.7.2>
```

Prepare to start Hadoop

Edit the file `etc/hadoop/hadoop-env.sh` and set JAVA_HOME:

```
# set to the root of your Java installation
  export JAVA_HOME=<the location of java on your machine>
```

For example, on my system, I set:

```
export JAVA_HOME=/usr/lib64/jvm/java-1.8.0-openjdk-1.8.0/
```

Test by typing

$ /bin/hadoop

Setup password-less SSH

Check that you can ssh to the localhost without a passphrase:

```
$ ssh localhost
```

If you cannot ssh to localhost without a passphrase, execute the following commands:

```
$ ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

Configure HDFS

Edit **`etc/hadoop/core-site.xml`** and add:

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Edit **`etc/hadoop/hdfs-site.xml`** and add:

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

Setup and start HDFS

Work in your Hadoop distribution:

1. Format the filesystem:

   ```
   $ bin/hdfs namenode -format
   ```

2. Start the NameNode and DataNode daemons:

   ```
   $ sbin/start-dfs.sh
   ```

   Note: you now have log output in the logs directory.

3. Open a browser and check the NameNode at:
   - http://localhost:50070/

4. Make the HDFS directories required to execute MapReduce jobs:

   ```
   $ bin/hdfs dfs -mkdir /user
   $ bin/hdfs dfs -mkdir /user/<username>
   ```

Configure YARN and MapReduce

Edit **`etc/hadoop/mapred-site.xml`** and add:

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

Edit **`etc/hadoop/yarn-site.xml`** and add:

```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>
```

Start the YARN daemons

- Start ResourceManager daemon and NodeManager daemon:

  $ sbin/start-yarn.sh

- Browse the web interface for the ResourceManager at:
  - http://localhost:8088/

Test by running a MapReduce job

Example - run the Distributed Shell app:

```
bin/hadoop \
jar share/hadoop/yarn/hadoop-yarn-applications-distributedshell-2.7.2.jar \
org.apache.hadoop.yarn.applications.distributedshell.Client \
--jar share/hadoop/yarn/hadoop-yarn-applications-distributedshell-2.7.2.jar \
--shell_command date --num_containers 2 --master_memory 1024
```

Windows installation

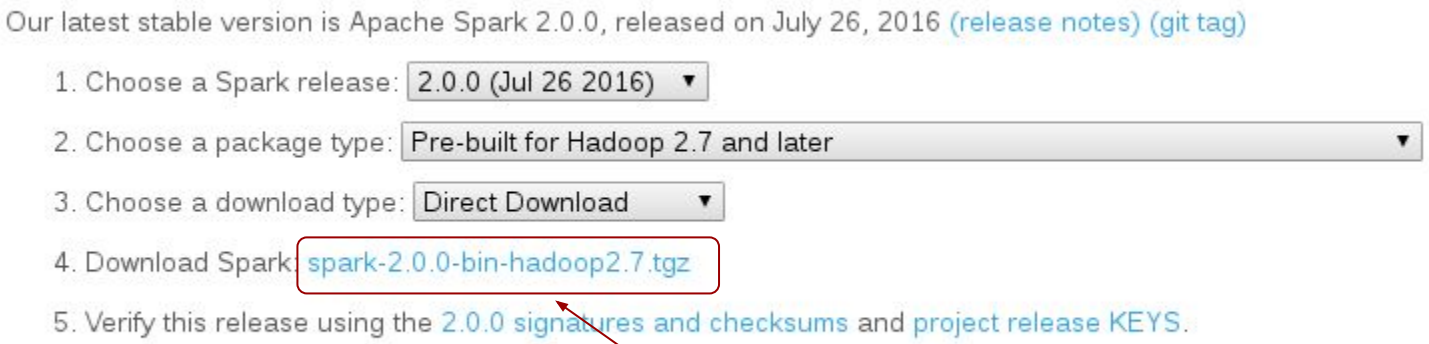
see:

http://wiki.apache.org/hadoop/Hadoop2OnWindows

Download Spark for Hadoop 2.7 and above

Download Spark from here.

I chose these options:



The download is 178 MG - will take a few minutes

Untar Spark and run a few checks

Untar Spark:

```
tar xvf spark-2.0.0-bin-hadoop2.7.tgz
```

Change directories into the Spark distribution:

```
spark-2.0.0-bin-hadoop2.7
```

Run an example here:

```
bin/run-example SparkPi 10
```

Perform a few more checks provided here:

http://spark.apache.org/docs/latest/#running-the-examples-and-shell