Spark环境搭建文档（可不依赖hadoop，1.6.2版本Spark）：

Spark环境依赖于Java和Scala。其中Java安装不再赘述，需要JRE 1.7+。

1. Scala下载：

Spark需要scala环境，http://www.scala-lang.org/download/2.10.6.html
选择2.10.6版本安装。Spark 1.6.2需要Scala 2.10版本的支持，不兼容2.11+。
解压下载包到指定路径:

```
tar -zxvf scala-2.10.6.tgz /to/the/scala_path
```

2. Spark 下载：

进入http://spark.apache.org/downloads.html 官网，Spark release选择1.6.2，package type选择Pre-built for Hadoop 2.6, 下载spark-1.6.2-bin-hadoop2.6.tgz。
解压spark-1.6.2-bin-hadoop2.6.tgz至指定路径:

```
tar -zxvf spark-1.6.2-bin-hadoop2.6.tgz /to/the/spark_path
```

3. 环境变量：

环境变量可写在/etc/profile中，也可以写在/.bashrc等配置文件中,写好后使用source命令使其生效。

```
vim /etc/profile  #或者 vim ~/.bashrc
```

环境变量中的添加内容如下：

```
#Scala Config
export SCALA_HOME=/to/the/scala_path   #scala解压目录
export PATH=$PATH:$SCALA_HOME/bin

#Spark Config
#export MAVEN_OPTS="-Xmx3g -XX:MaxPermSize=1g -XX:ReservedCodeCacheSize=1g" #可选的maven开发选项
export SPARK_HOME=/to/the/spark_path  #Spark解压目录
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
```

环境变量添加完毕后，使其生效。注意，/etc/profile中的环境变量需要重启机器才能每次启动终端都生效，/.bashrc环境变量每次启动终端都直接生效。

```
source /etc/profile  #或者 source ~/.bashrc
```

4. Spark配置

Spark的配置文件在${SPARK_HOME}/conf目录下。

```
cd $SPARK_HOME
cp spark-env.sh.template spark-env.sh
cp log4j.properties.template log4j.properties
cp spark-defaults.conf.template spark-defaults.conf
cp slaves.template slaves
```

- spark-env.sh配置需添加:

```
export JAVA_HOME=${JAVA_HOME}
export SPARK_MASTER_IP=${your_master_ip}  # master ip
export SPARK_MASTER_PORT=7077
export SPARK_WORKER_CORES=1     #worker占用核数
export SPARK_WORKER_INSTANCES=2 #worker启动实例个数
export SPARK_WORKER_MEMORY=2g   #worker运行内存
export SPARK_WORKER_WEBUI_PORT=8081
export SPARK_EXECUTOR_CORES=1   #master调度器核数
export SPARK_EXECUTOR_MEMORY=1g   #master调度器内存
export SPARK_LOCAL_IP=${your_local_ip}  # work运行机 ip
export SPARK_SCALA_VERSION="2.10.6"
```

- spark-defaults.conf配置需添加:

```
spark.master                     spark://${master_hostname}:7077
spark.eventLog.enabled           true
spark.eventLog.dir               file:///to/the/spark/eventLogs/
#spark.eventLog.dir              hdfs://hostname:9000/spark-events
spark.serializer                 org.apache.spark.serializer.KryoSerializer  #默认序列化
spark.driver.memory              1g   #调度器驱动
spark.executor.memory            1g   #调度器运行内存
```

- log4j.properties默认配置就好，需要记录在文件里再改
- slaves文件: slaves 跟hadoop启动一致，集群中每个worker机节点hostname都需要保存在slaves文件中。
  注意，**不包括**master的hostname，如果master机也需要启动worker，则master的hostname也可写入slaves文件中。

5. Spark集群启动方式(在此只介绍不依赖于hadoop的集群的standalone启动方式):

- [x] 在master机:

```
start-master.sh
```

即可启动。

- [x] 在slave机:

```
#start-slave.sh <master-spark-URL>
#如下，这里假设master机的hostname为master
start-slave.sh spark://master:7077
```

另注，在master机也可以启动slave，slave中是worker的工作环境。所有spark任务均执行在work中。
仅供参考。