一、安装Spark集群
1. 解压安装包
2. 配置spark
3. 复制到其他节点
4. 设置环境变量
5. 启动验证
6.打包运行
二,高可用(热备)

注:配置环境时,$SPARK_HOME/sbin一定放在hadoop的sbin前面,因为这两个文件夹中都含有start-all.sh和stop-all.sh,而spark启动/关闭用到了这两个脚本,而hadoop中这么命令已经遗弃了

> 源笔记:
> **1.修改slave文件,配置从机**
> 如下:
> slave01
> slave02
> slave03
> **2.修改spark.env.sh,如下:**
> ```
> export JAVA_HOME=/mysoftware/jdk1.8.0_101
> export SPARK_MASTER_IP=master
> export SCALA_HOME=/mysoftware/scala-2.12.3
> export HADOOP_HOME=/mysoftware/hadoop-2.7.3
> export HADOOP_CONF_DIR=/mysoftware/hadoop-2.7.3/etc/hadoop/
> ```
>
> **3.发送到从机即可,集群环境搭建完毕**

一、安装Spark集群

1. 解压安装包

tar -zxvf ~/jar/spark-1.6.3-bin-hadoop2.6.tgz -C /data

2. 配置spark

> 涉及到的配置文件有以下几个:
> `${SPARK_HOME}/conf/spark-env.sh`
> `${SPARK_HOME}/conf/slaves`
> `${SPARK_HOME}/conf/spark-defaults.conf`

这三个文件都是由原始的template文件复制过来的，比如cp spark-env.sh.template spark-env.sh

配置文件1：spark-env.sh

```
JAVA_HOME=/data/jdk1.8.0_111
SCALA_HOME=/data/scala-2.11.8
SPARK_MASTER_HOST=master
SPARK_MASTER_PORT=7077
HADOOP_CONF_DIR=/data/hadoop-2.6.5/etc/hadoop
# shuffled以及RDD的数据存放目录
SPARK_LOCAL_DIRS=/data/spark_data
# worker端进程的工作目录
SPARK_WORKER_DIR=/data/spark_data/spark_works
```

> 注意：需要在本地创建/data/spark_data/spark_works目录

配置文件2：slaves

```
master
slave1
slave2
```

配置文件3：spark-defaults.conf

```
spark.master        spark://master:7077
spark.serializer        org.apache.spark.serializer.KryoSerializer
spark.eventLog.enabled        true
```

spark.eventLog.dir file:///data/spark_data/history/event-log
spark.history.fs.logDirectory file:///data/spark_data/history/spark-events
spark.eventLog.compress true

> 注意：需要在本地创建/data/spark_data/history/event-log、/data/spark_data/history/spark-events

3. 复制到其他节点

在master上：

```
scp -r /data/spark* aboutyun@slave1:~/
scp -r /data/spark* aboutyun@slave2:~/
```

在slave1和slave2上：

```
mv ~/spark* /data
```

4. 设置环境变量

将以下内容加入到~/.bashrc文件中，

```
export SPARK_HOME=/data/spark-1.6.3-bin-hadoop2.6
export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH
```

然后执行以下命令：

```
source ~/.bashrc
```

5. 启动验证

在master机器上进行如下操作

> 可直接使用 start-all.sh ，但是要注意hadoop的命令, hadoop也有一个一样的命令,但是已经遗弃

1）启动master

```
start-master.sh
```

在master机器上执行jps命令

```
[aboutyun@master conf]$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /data/spark-1.6.3-bin
-hadoop2.6/logs/spark-aboutyun-org.apache.spark.deploy.master.Master-1-master.ou
t
[aboutyun@master conf]$ jps
6371 Master
3367 NameNode
3850 ResourceManager
6428 Jps
3598 SecondaryNameNode
[aboutyun@master conf]$
```

上图说明在master节点上成功启动Master进程

2）启动slave

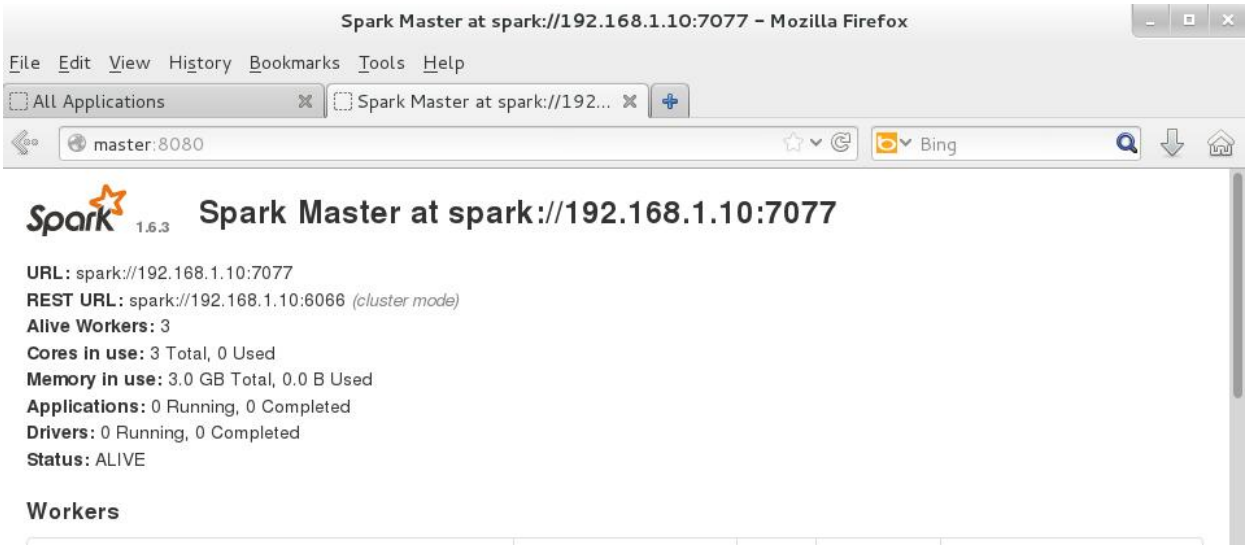
在master和slave机器上执行jps命令

```
[aboutyun@master conf]$ start-slaves.sh
slave1: starting org.apache.spark.deploy.worker.Worker, logging to /data/spark-1
.6.3-bin-hadoop2.6/logs/spark-aboutyun-org.apache.spark.deploy.worker.Worker-1-s
lave1.out
slave2: starting org.apache.spark.deploy.worker.Worker, logging to /data/spark-1
.6.3-bin-hadoop2.6/logs/spark-aboutyun-org.apache.spark.deploy.worker.Worker-1-s
lave2.out
master: starting org.apache.spark.deploy.worker.Worker, logging to /data/spark-1
.6.3-bin-hadoop2.6/logs/spark-aboutyun-org.apache.spark.deploy.worker.Worker-1-m
aster.out
[aboutyun@master conf]$ jps
7282 Jps
6371 Master
3367 NameNode
7208 Worker
3850 ResourceManager
3598 SecondaryNameNode
[aboutyun@master conf]$
```

上面的图片说明在每台机器上都成功启动了Worker进程。

3）访问WebUI

在master、slave1和slave2这三台中任意一台机器上的浏览器中输入：
http://master:8080/，看到如下图片，就说明我们的spark集群安装成功了。



趟过的坑

1. 配置core-site.xml和hdfs-site.xml文件时所指定的本地目录一定要自己创建，否则在执行玩格式化hdfs后，启动hdfs会丢失进程。

6.打包运行

https://www.cnblogs.com/654wangzai321/p/9513488.html

二,高可用(热备)