采用最普遍的Spark on Yarn模式，该模式前景最广阔，企业里用的最广。

## 软件下载和上传

Scala linux版本：[http://downloads.lightbend.com/scala/2.11.8/](http://downloads.lightbend.com/scala/2.11.8/scala-2.11.8.tgz)[scala-2.11.8.tgz](http://downloads.lightbend.com/scala/2.11.8/scala-2.11.8.tgz)

Spark前边已下载过：spark-2.2.0-bin-hadoop2.6.tgz

## 部署

我们的CDH中用**CM**进行集群管理，集群直接互联是通过ssh协议，但我们不需要配置ssh免密匙访问，因为CM中配置了通过相同帐户密码访问。

用Apache Spark的话，必须配置ssh免密匙访问。

### 配置免密匙访问

设置ssh无密码访问

1、执行ssh-keygen -t rsa

#建立ssh目录,一路敲回车，生成的密钥对id\_rsa，id\_rsa.pub，

默认存储在~/.ssh目录下

chmod 755 .ssh #赋予755权限

cd .ssh

#ls –l

id\_rsa id\_rsa.pub

2、cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

#把公用密匙添加到authorized\_keys 文件中（此文件最后一定要赋予644权限）

ssh slave1 cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

#上面条命令是把其他slave节点的公钥拷贝到master节点中的authorized\_keys，

#有几个slave节点就需要运行几次命令，slave1是节点名称

scp ~/.ssh/authorized\_keys slave1:~/.ssh/

#把authorized\_keys文件拷贝回每一个节点，slave1是节点名称

最后通过ssh slave1 测试下。

### 解压Scala和Spark

1. 删除cdh中的Spark：

rm -rf /usr/bin/spark\*

rm -rf /etc/spark

1. 上传至spark-2.2.0-bin-hadoop2.6 和 scala-2.11.8.tgz 至 /opt/soft/spark2.0下，并进行解压：

tar xvzf spark-2.0.0-preview-bin-hadoop2.6.tgz

tar xvzf scala-2.11.8.tgz

vi /etc/profile，增加如下内容：

export HADOOP\_HOME=/opt/cloudera/parcels/CDH/lib/hadoop

export SPARK\_HOME=/opt/soft/spark2.0/spark-2.2.0-bin-hadoop2.6

export SCALA\_HOME=/opt/soft/spark2.0/scala-2.11.8

export JAVA\_HOME=**/opt/soft/jdk1.8.0\_131**

export HADOOP\_CONF\_DIR=/etc/hadoop/conf

source /etc/profile 起效

1. 修改SPARK\_HOME/conf下

mv slaves.template slaves ，slaves 里配置工作节点主机名列表

mv spark-env.sh.template spark-env.sh ，spark-env.sh配置一些环境变量，由于我们用Yarn模式，这里面不用配置。

1. 运行测试

在2.0之前，Spark在YARN中有yarn-cluster和yarn-client两种运行模式，建议前者。

而在2.0 里--master的yarn-cluster 和 yarn-client 都deprecated了，统一用yarn 。

用run-example 方便测试环境：

run-example SparkPi local模式运行

分布式模式运行：

spark-submit --class org.apache.spark.examples.SparkPi \

--master yarn \

--num-executors 1 \

--driver-memory 1g \

--executor-memory 1g \

--executor-cores 1 \

--conf "spark.app.name=SparkPi" \

/opt/soft/spark2.0/spark-2.2.0-bin-hadoop2.6/examples/jars/spark-examples\_2.11-2.2.0.jar

测试spark-shell ，直接执行spark-shell即可。

**官方提供的examples在tgz中，是学习Spark最直接有效的方式！**

**大家要把例子包导入IDE中进行学习。**

**如果内存不足，报错的话，在cm里进行yarn 的配置**，如下2个设置为2g:

yarn.scheduler.maximum-allocation-mb

yarn.nodemanager.resource.memory-mb

保存后进行：



部署客户端配置的作用：把cm界面里修改过的参数同步到每个节点的xml配置文件里。

然后重启Yarn服务起效。