# 环境条件

## CDH

访问地址：<http://bigdata5:7180/cmf/home>

CDH服务端位于 bigdata5节点，而客户端分别位于bigdata1、bigdata3、bigdata4节点。

目前已有服务如图 1所示：



图 1 CDH现存服务

在当前环境中，已经有Spark服务，该服务版本为Spark1.6.3，本文档的目标是添加Spark2服务和Spark1共存。

观察Spark1的主机角色配置，如图 2所示：



图 2 Spark服务的主机角色配置

在图 2中可以看到四个节点都作为了Gateway，而bigdata4作为了History Server，同理为了保持一致性，本文档也采用这种主机角色配置方案。

## 集群



图 3 集群节点详细信息

如图 3所示总共四个主机，每个主机都具有相应的角色。由于对于CDH服务而言，其服务端位于bigdata5，因此我们这里主要的部署方案是在bigdata5上进行，然后进行资源分发到各个客户端节点。

## 软件包

Spark2的软件包的版本主要被CDH版本所限制，观察CDH版本，如图 4所示，可知CDH为5.8版本。

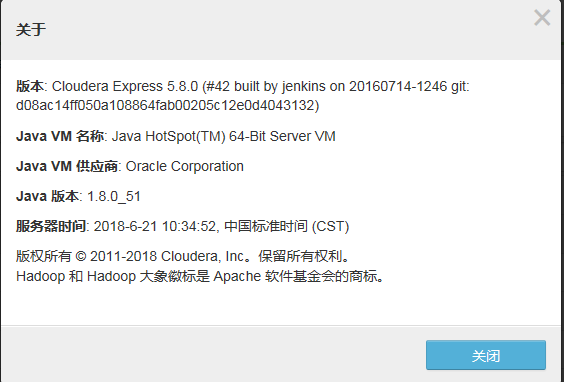


图 4 CDH版本信息

当前Spark2软件包可以从[cloudera官网](https://www.cloudera.com/documentation/spark2/latest/topics/spark2_packaging.html#versions)查看，如图 5所示。

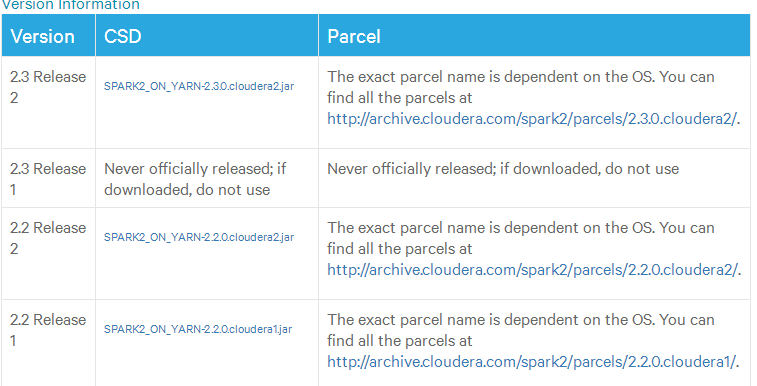
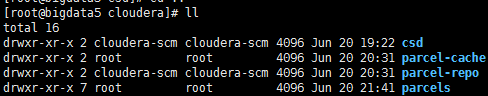


图 5 Spark2软件包

对于Spark2来说，原则上是使用最新版本的软件包，也就是最新版的2.3，但是在尝试使用后，发现必须将CDH升级到5.9版本才行，因此本文档采用2.2版本，也即图 5中对应2.2 Release 2。下载“SPARK2\_ON\_YARN-2.2.0.cloudera2.jar”到bigdata5服务器的/opt/cloudera/csd目录，并修改所属用户和权限。



所属用户和csd一致，用户和组都是cloudera-scm

|  |
| --- |
| cd /opt/cloudera/csd  chown cloudera-scm:cloudera-scm SPARK2\_ON\_YARN-2.2.0.cloudera2.jar  chmod 644 SPARK2\_ON\_YARN-2.2.0.cloudera2.jar |



重启sever：

|  |
| --- |
| service cloudera-scm-server restart |

从Parcel源的链接中下载三种类型的文件。下载时，主要涉及到两种种版本的对应：

1. CDH版本，我们集群是CDH5.8，因此这里对应cdh5；
2. Linux系统版本，我们的系统是CentOS6，因此对应el6;

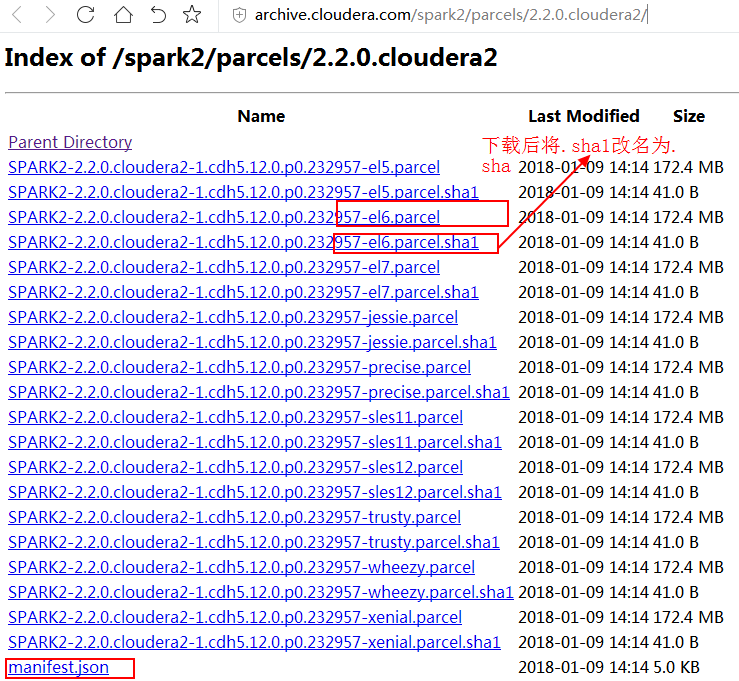
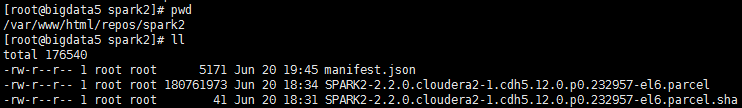


图 6 parcels源要下载的文件

在图 6中，要注意的是.sha1文件下载后要修改后缀为.sha，否则CDH可能无法识别该源。另外一种操作是直接将该网络地址复制到CDH UI中让CDH自动远程下载，但是经过测试发现，CDH自动下载耗时太久，因此本文档采用本地导入源的方式。

当下载完成后，到bigdata5服务器上进行以下操作：

|  |
| --- |
| # 进入到CDH的远程源仓库中  cd /var/www/html/repos/  # 创建spark2文件夹  mkdir spark2  # 将上面下载的三个文件都发送到spark2目录中 |



至此，软件包下载成功。

# Spark2部署

## 配置Parcel

进入到CDH UI界面，如图 7所示，在主界面的右上角点击Parcel图标，然后在弹出的界面中点击配置按钮。



图 7 配置Parcel步骤

现在你可以看到集群中的所有的Parcel源信息，然后把放在/var/www/html/repos/目录下的Parcel配置进去，如图 8所示。

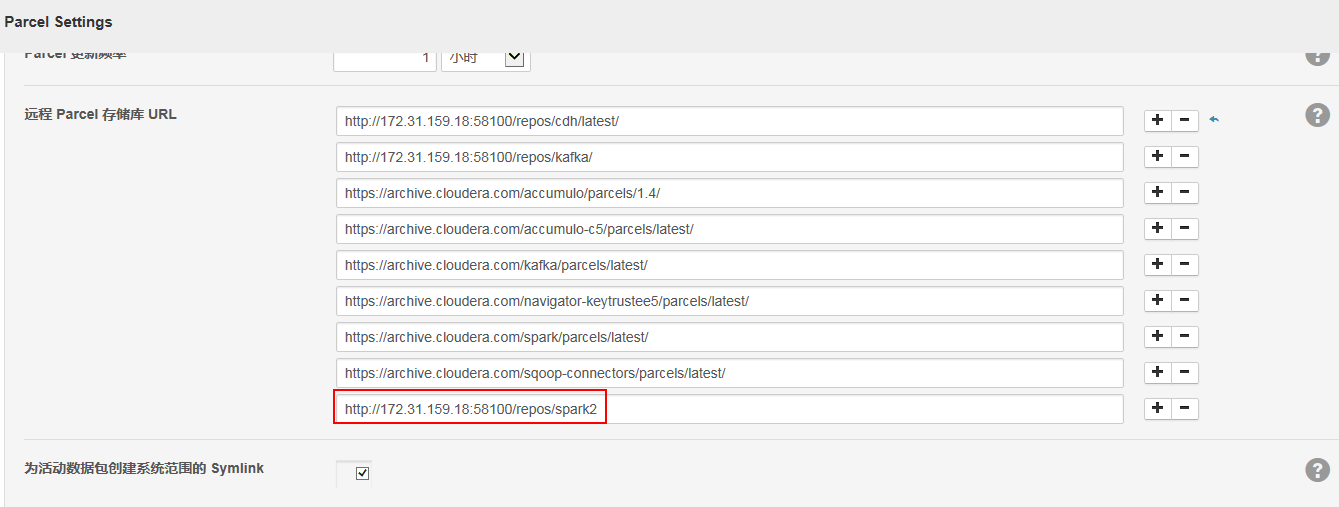


图 8 配置Parcel源的库位置

然后刷新页面能够发现在Parcel中自动出现SPARK2信息，如图 9所示。



图 9 Pacel列表

现在逐步点击SPARK2的下载，分配，激活，操作完成后，其状态变为“已分配，已激活”。至此，表示Spark2的源已经分配到各个节点，且采用了指定的软件包版本。

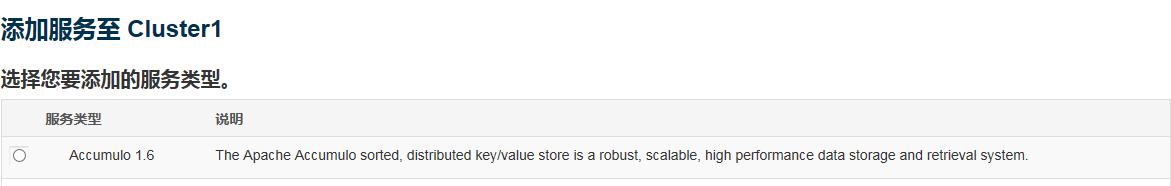
## 安装Spark2服务

回到CDH主页，如图 10所示，点击添加服务链接。



图 10 为CDH添加新的服务

选择Spark2服务，如图 11所示



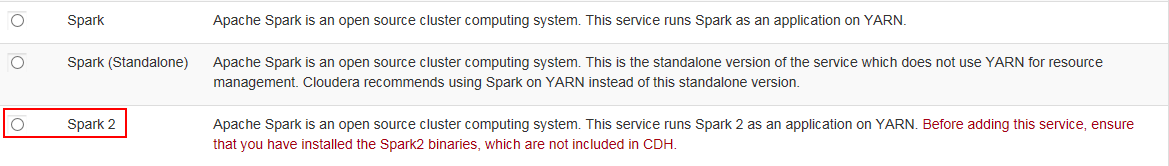


图 11 选择Spark2服务

选择能够支持Hive的Spark2版本，如图 12所示。

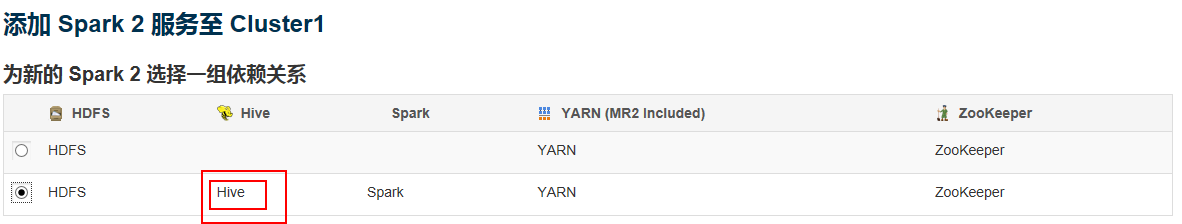


图 12 选择支持Hive的Spark2版本

角色分配中考虑和Spark1的设置保持一致，如图 12所示。History Server选择bigdata4，而Gateway则是将四个主机都选择。



图 13 自定义Spark2的角色分配

审核功能选择默认配置即可，如图 14所示

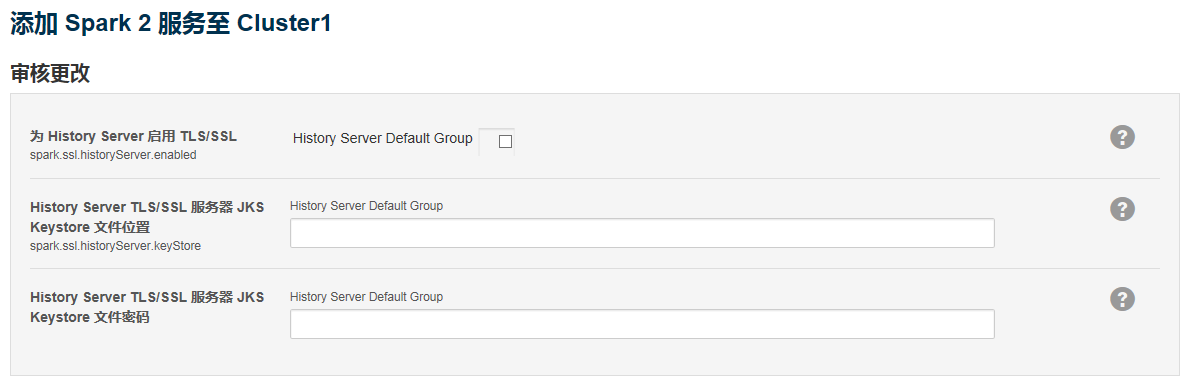


图 14 审核功能选择默认配置

然后等待完成。就表示添加服务步骤结束。回到主页，查看服务列表。如图 15所示，可以看到Spark2已经添加成功了。



图 15 添加服务之后的集群

最后任意提交一个Spark代码采用spark2-submit命令来提交到集群。代码举例：

|  |
| --- |
| **object** put {  **def** main(args: Array[String]): Unit = {**val** spark = SparkSession.*builder*().getOrCreate()  spark.sparkContext.setLogLevel(**"WARN"**)  **val** cut = spark.sql(**"show databases"**).count()  **while**(**true**){  *println*(**"$$$$$$$$%d$$$$$$$$"**.format(cut))  Thread.*sleep*(10000)  }  spark.stop()  } } |

提交命令为：

|  |
| --- |
| spark2-submit --class com.ruijie.stream.put --name pengxx-test-spark2 --master yarn --deploy-mode cluster /home/pengxx/spark/spark2-test-dev.jar |

运行结果能够在Yarn集群中看到，则说明Spark2服务部署成功。

