# Hive on Spark实施笔记

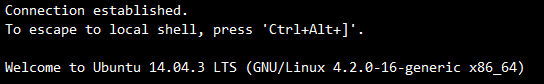
## 编译适合Hive的Spark

## 说明

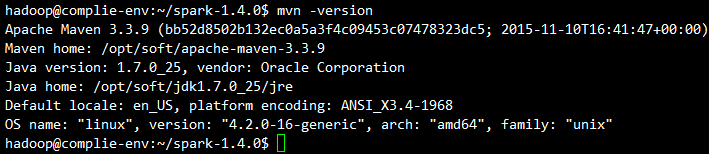
Spark为了支持从Hive中读取数据，所以有很多Hive中用到的jar包，而Hive中Hive on spark时会将Spark的jar包引入到Hive的运行环境，因此及其他原因会有jar冲突。所以需要重新编译没有包含Hive相关模块的Spark。

## 编译环境

* + 1. OS – Ubuntu 14.04



* + 1. Maven



* + 1. Scala (Spark编译时有用到)



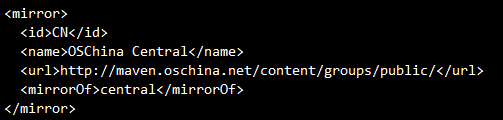
<http://www.scala-lang.org/files/archive/nightly/2.11.x/scala-2.11.7-950bb26-nightly.tgz>

* + 1. Hadoop2.6(已引入HADOOP\_HOME环境变量)
    2. 可访问外网（编译过程要联网下载）
    3. 配置maven的国内镜像（编译过程会从国外下载较多文件，改国内maven库镜像后编译过程大约需要1小时）

设置本地库路径：



镜像：



## 编译

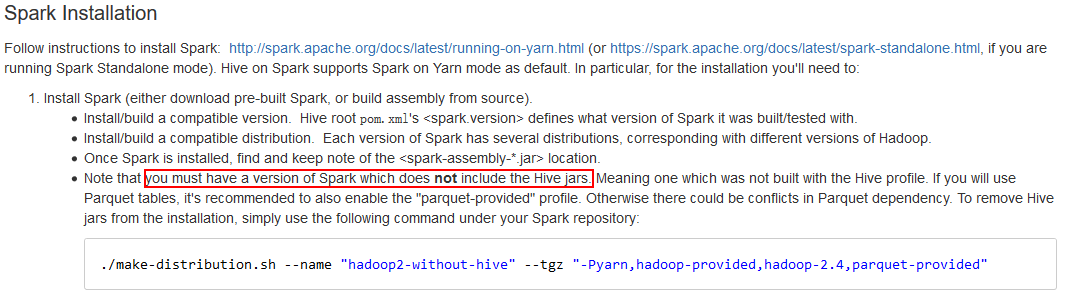
* + 1. 下载spark 源码

为了避免hive可能不支持到spark新版，以及网上信息反映和投入时间，选择较稳定的spark-1.4.0.tgz

<http://spark.apache.org/releases/spark-release-1-4-0.html>

* + 1. 编译

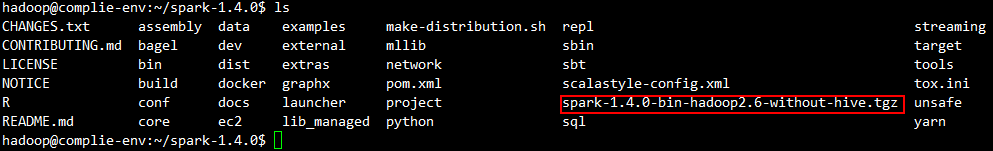
建议按照hive on spark官网的spark编译方式编译：



我用的编译命令是:



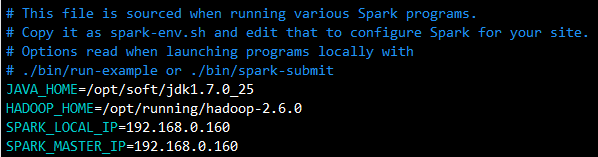
编译完成后得到最终文件：



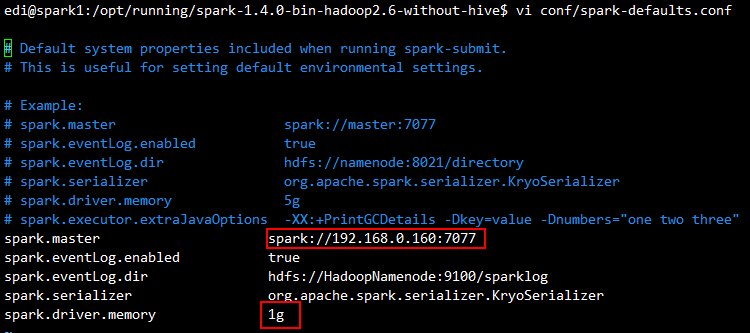
* + 1. 安装

复制结果文件到部署机子的安装目录用tar zxvf解压，配置SQOOP\_HOME并将bin加入到PATH。部署集群需要中集群各节点安装spark。

* + 1. 配置
       1. **conf/spark-env.sh**



* + - 1. **conf/spark-defaults.conf**



* + - 1. conf/slaves

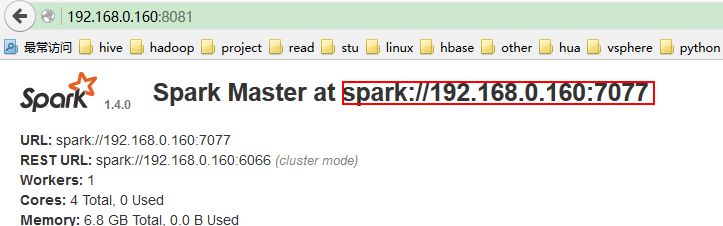
加入spark的slave节点IP（一行一个）。

* + - 1. conf/log4j.properties 如果有问题可通过该文件配置log级别，输出明细。
    1. 启动

运行环境：jdk1.7，hadoop2.6，hvie1.2.1

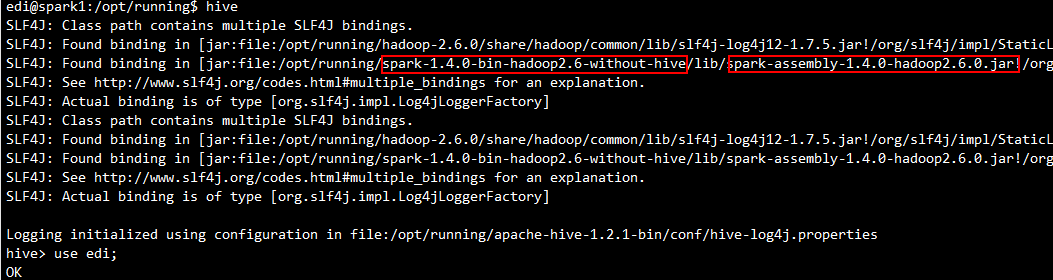
启动spark集群服务:/sbin/start-all.sh ，启动失败可能是编译有问题。

启动后可以从webUI查看Spark Master URL

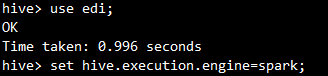


## 配置Hive On Spark

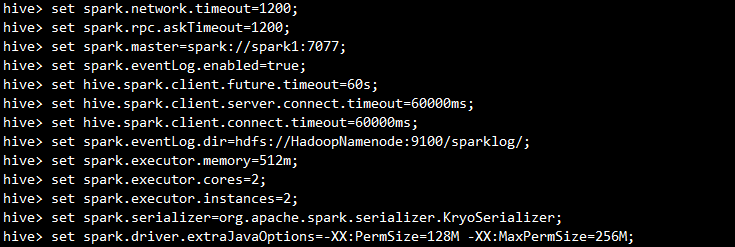
* + 1. Hive启动时会主动检查是否有Spark可用，如果上面有正确配置过SPARK\_HOME环境变量则Hive会引入Spark以便支持Hive on spark .正常启动输出如下:



* + 1. 然后Hive根据配置参数hive.execution.engine 确定接下来计算使用的计算框架，该参数默认是mr，支持tez\spark。



Hive on spark常用的参数还有：

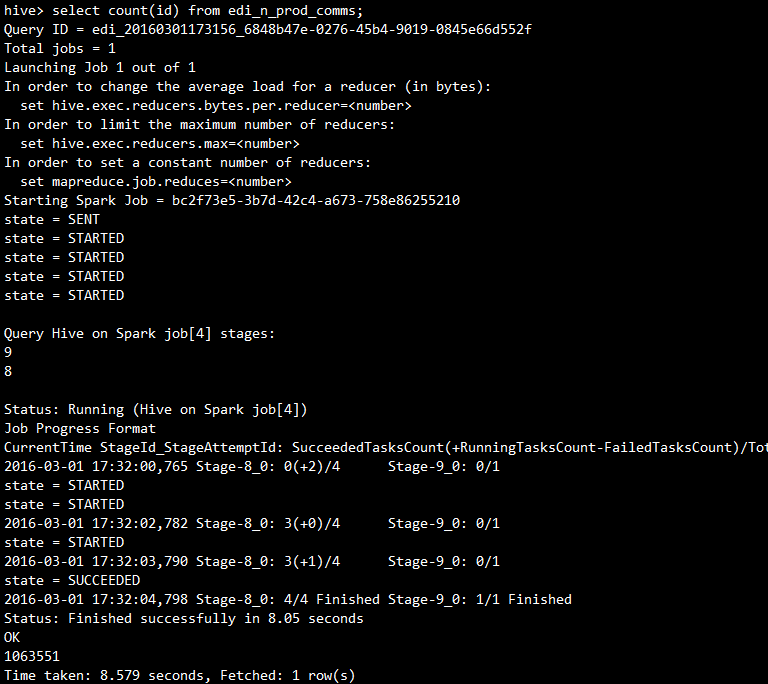


## 测试

执行会触发聚合计算的hive SQL ,如：



集成正常的日志如下：



## 参考信息

* + 1. <https://cwiki.apache.org/confluence/display/Hive/Hive+on+Spark:+Getting+Started#HiveonSpark:GettingStarted-ConfiguringHive>
    2. <http://spark.apache.org/docs/latest/running-on-yarn.html>
    3. <http://blog.csdn.net/wind520/article/details/46762173>
    4. <http://www.iteblog.com/archives/1541>