spark--------编译源码

2016/11/30

裴卫

目录

Spark--------编译源码 1
使用系统： 3
安装环境 3
准备 3
安装 maven 3
检查 3
配置 3
源 4
编译 4
源码 4
build/mvn 方法 5
Building a Runnable Distribution 方法 5
监控 5

使用系统：

Ubuntu16.04
配有 LAMP，vim
系统配置 2G 内存 60G 硬盘 配置还是比较差

安装环境

Jdk：1.8
Maven:

准备

Spark 源码：spark2.0.2
安装 jdk 太简单，此处略去（spark 搭建文档中有具体信息）

安装 maven

网上文档太多，良莠不齐，不对他人报告作评价

检查

首先要检查 jdk 的安装情况，因为 maven 依赖于 jdk
java –version
先安装：apt-get install maven
安装后文件位置：/usr/share/maven

配置

设置环境变量：sudo vim /etc/profile
在最后面加以下内容：
Jdk 1.7 及以下
M2_HOME=/usr/local/apache-maven-3.1.0
export MAVEN_OPTS="-Xms256m -Xmx512m"
export PATH=$M2_HOME/bin:$PATH

jdk 1.7 以上
export M2_HOME=/usr/share/maven
export M2=$M2_HOME/bin
export PATH=$M2:$PATH

执行 source /etc/profile 使环境变量生效
检查安装：mvn –version
看安装位置以及版本

源

由于源的问题，我们使用的是阿里云的源
阿里云 nexus 地址 http://maven.aliyun.com/nexus/#welcome

进入/usr/local/maven 文件夹，在 conf 目录中找到 settings.xml 文件，配置 mirrors 的子节点，添加如下 mirror
阿里云 Maven 镜像：

```
<mirror>
      <id>nexus-aliyun</id>
      <mirrorOf>central</mirrorOf>
      <name>Nexus aliyun</name>
      <url>http://maven.aliyun.com/nexus/content/groups/public</url>
</mirror>
```

或开源中国 maven 镜像（开源中国镜像好像封了）

```
<mirror>
   <id>nexus-osc</id>
   <mirrorOf>*</mirrorOf>
   <name>Nexus osc</name>
   <url>http://maven.oschina.net/content/groups/public/</url>
</mirror>
```

编译

源码

Spark 源码地址在 github 上有

https://github.com/apache/spark
我是在计算机上下载下来，再上传到 Ubuntu 集群上进行编译的
Spark 官网提供两种编译方式的讲解以及支持，有兴趣可以看官网
http://spark.apache.org/docs/latest/building-spark.html

build/mvn 方法

进入 spark 源码目录，看是否有 pom,xml 文件，若有使用，没有进 build 文件夹里面使用此命令
mvn -Phadoop-2.7 -Pyarn -DskipTests -Dhadoop.version=2.7.0 -Pspark-ganglia-lgpl clean package
(这种方式，推荐，亲测有用)

Building a Runnable Distribution 方法

进入 spark 源码目录，再进入 Dev 文件夹，带参数运行 make-distribution.sh

./make-distribution.sh --tgz --name 2.2.0 -Pyarn -Phadoop-2.2 -Pspark-ganglia-lgpl -Phive
格式： ./make-distribution.sh [--name] [--tgz] [--with-tachyon] <maven build options>

- --with-tachyon：是否支持内存文件系统 Tachyon，不加此参数时不支持 tachyon。
- --tgz：在根目录下生成 spark-$VERSION-bin.tgz，不加此参数时不生成 tgz 文件，只生成/dist 目录。
- --name NAME：和--tgz 结合可以生成 spark-$VERSION-bin-$NAME.tgz 的部署包，不加此参数时 NAME 为 hadoop 的版本号。

如果要生成 spark 支持 yarn、hadoop2.2.0、hive 的部署包，只需要将源代码复制到指定目录，进入该目录后运行：
./make-distribution.sh --tgz --name 2.2.0 -Pyarn -Phadoop-2.2 -Phive
如果要生成 spark 支持 yarn、hadoop2.2.0、ganglia、hive 的部署包，只需要将源代码复制到指定目录，进入该目录后运行：
./make-distribution.sh --tgz --name 2.2.0 -Pyarn -Phadoop-2.2 -Pspark-ganglia-lgpl -Phive

数字为 Hadoop 版本号，此方法未亲测，慎用

监控

先进入 spark 编译后的文件夹中，再进入 conf 文件夹，将 metrics.properties.template 文件复制一份命名为 metrics.properties

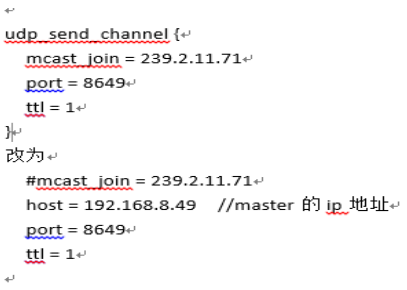
将文件最后几行注释去掉
原文件最后几行

Enable JvmSource for instance master, worker, driver and executor
#master.source.jvm.class=org.apache.spark.metrics.source.JvmSource
#worker.source.jvm.class=org.apache.spark.metrics.source.JvmSource
#driver.source.jvm.class=org.apache.spark.metrics.source.JvmSource
#executor.source.jvm.class=org.apache.spark.metrics.source.JvmSource

改为
# Enable JvmSource for instance master, worker, driver and executor
master.source.jvm.class=org.apache.spark.metrics.source.JvmSource
worker.source.jvm.class=org.apache.spark.metrics.source.JvmSource
driver.source.jvm.class=org.apache.spark.metrics.source.JvmSource
executor.source.jvm.class=org.apache.spark.metrics.source.JvmSource

Vim metrics.properties 在最后面将下面内容贴上去
*.sink.ganglia.class=org.apache.spark.metrics.sink.GangliaSink
*.sink.ganglia.host=192.168.1.105 //填主机 ip
*.sink.ganglia.port=8649 //端口要与 ganglia 设置的一样
*.sink.ganglia.period=10
*.sink.ganglia.unit=seconds
*.sink.ganglia.ttl=1
*.sink.ganglia.mode=multicast

附上 ganglia 配置信息



在 master 节点上通过下面方法重启服务。

sudo /etc/init.d/ganglia-monitor start

sudo /etc/init.d/gmetad start

sudo /etc/init.d/apache2 restart

最后执行

sudo /etc/init.d/ganglia-monitor restart

在刷新监控页面，查看资源是不是加入到监控中。
进入监控 web 进行查看，应该会有 worker，jvm 等信息