**LPW使用手册**

1. **简介**

为更好的选择Apache Spark框架运行时的缓存数据，本项目提出了LPW(Least partition weight) 模型，LPW是一个面向分布式计算框架Spark的缓存替换系统，可以为用户提供数据缓存替换功能。由于应用负载特征的多样性、内存资源需求的变化性以及内存资源的持续变化性，LPW根据数据块计算代价、数据块引用次数、数据块所占内存空间等因素建立模型，通过权重识别缓存数据的必要性，并与Spark无缝集成，在应用执行时提供更合理的数据进行替换，提升应用执行的效率。

1. **环境准备**

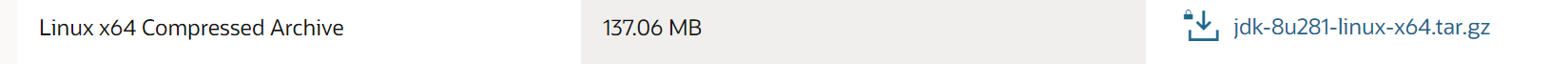
使用LPW需要搭建基础环境，包括Java、Scala、Maven组件的安装。同时，因为LPW是基于Spark扩展的，所以还需要下载spark2.2.3源码。通过安装的Maven对spark源码进行编译。后续脚本的执行和提交与原生spark的使用方式相同。

* 1. **安装JAVA**

下载地址：

<https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>

注：这里需要有oracle账户登录。



配置环境变量。

* 1. **安装Scala**

下载地址：<https://downloads.lightbend.com/scala/2.12.11/scala-2.12.11.tgz>

配置环境变量。

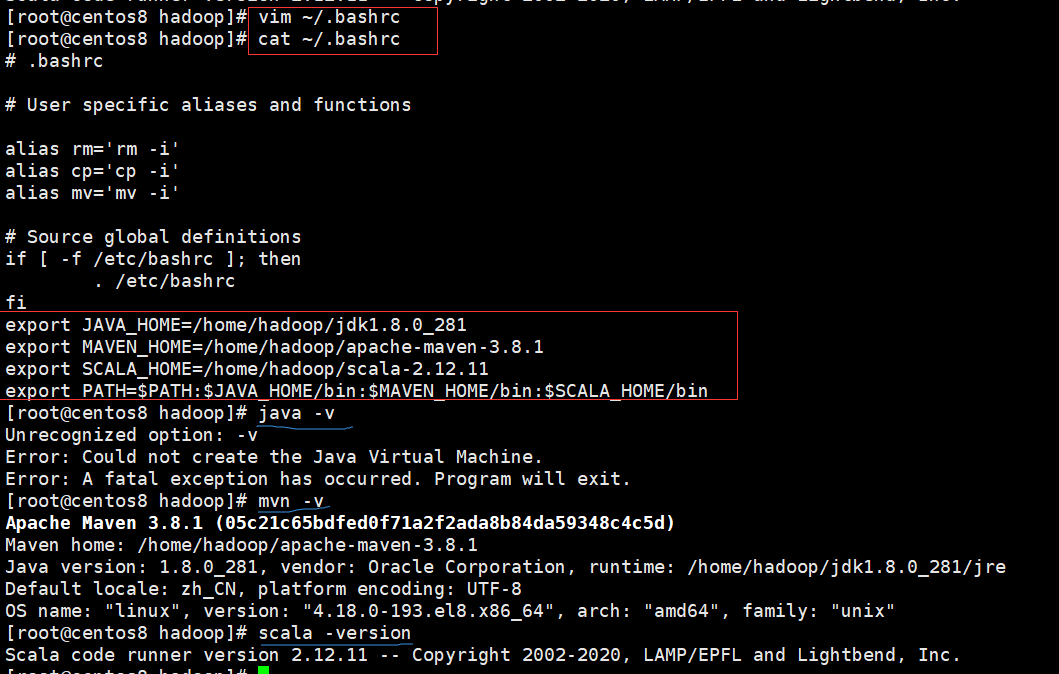
* 1. **安装MAVEN**

Maven下载地址：

<https://archive.apache.org/dist/maven/maven-3/3.6.2/binaries/>



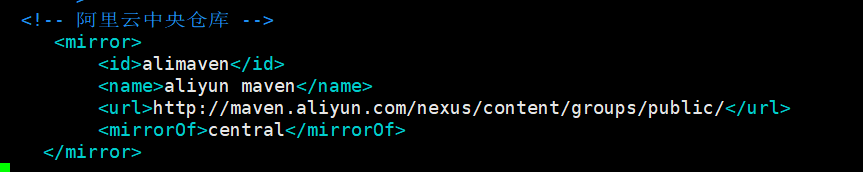
配置环境变量



环境变量配置后一定要保存并更新，**source ~/.bashrc**

* 配置阿里云中央仓库

/home/hadoop/apache-maven-3.8.1/conf/settings.xml 配置文件添加以下内容

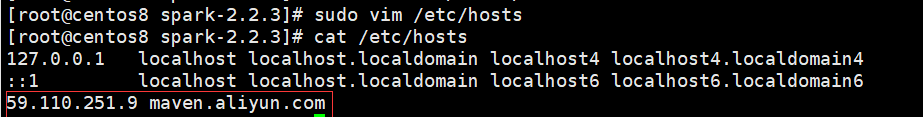


测试阿里云仓库是否能正确访问：



如果不能正确访问，解决办法如下：

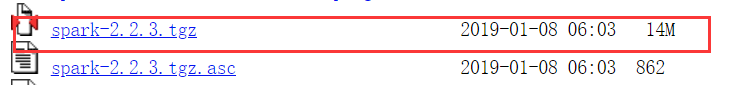
本机host配置maven.aliyun.com的域名，截图如下：



* 1. **编译Spark2.2.3源码**

下载源码：<https://archive.apache.org/dist/spark/spark-2.2.3/>

wget https://archive.apache.org/dist/spark/spark-2.2.3/spark-2.2.3.tgz



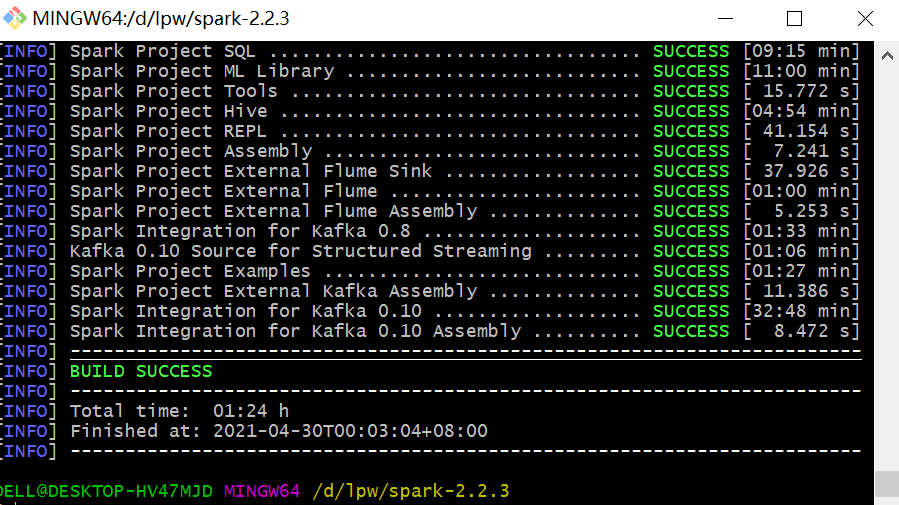
进入源码的根目录（pom.xml所在的目录）

* 编译命令

mvn -DskipTests clean package



* 验证编译是否成功



以上截图以windows平台为测试平台，linux平台同样适合。

1. **编译安装**

本项目的编译安装有两种途径：1、将项目提供的jar包放至指定路径。2、替换spark源码文件重新编译。下面讲解两种途径的操作方法。

* 1. **替换jar包**

本项目lpw-jar路径下提供有编译好的spark-core\_2.11-2.2.3.jar文件，将其拷贝到spark-2.2.3-bin-hadoop2.7/jars路径下覆盖原文件即可。

* 1. **替换源文件**

用本项目中Instrument文件夹下的scala文件分别替换原生spark中的同名文件。之后在spark根目录下使用mvn clean package -DskipTests进行编译。编译完成后即可将该修改过的spark，开启的lpw使用之旅。本工程中添加了许多便于测试及观察缓存替换效果的额外日志输出。用户可根据自己的需求将这些输出注释掉。

* + 1. **文件更新列表**

替换该工程目录下的文件，文件更新列表如下：

|  |  |
| --- | --- |
| **文件名** | **原文件位置** |
| MemoryStore.scala | core/src/main/scala/org/apache/spark/storage/memory/MemoryStore.scala |
| BlockId.scala | core/src/main/scala/org/apache/spark/storage/BlockId.scala |
| BlockManagerMaster.scala | core/src/main/scala/org/apache/spark/storage/BlockManagerMaster.scala |
| BlockManagerMasterEndpoint.scala | core/src/main/scala/org/apache/spark/storage/BlockManagerMasterEndpoint.scala |
| BlockManagerMessages.scala | core/src/main/scala/org/apache/spark/storage/BlockManagerMessages.scala |
| BlockManagerSlaveEndpoint.scala | core/src/main/scala/org/apache/spark/storage/BlockManagerSlaveEndpoint.scala |
| RDD.scala | core/src/main/scala/org/apache/spark/rdd/RDD.scala |
| DAGScheduler.scala | core/src/main/scala/org/apache/spark/scheduler/DAGScheduler.scala |
| ResultTask.scala | core/src/main/scala/org/apache/spark/scheduler/ResultTask.scala |
| ShuffleMapTask.scala | core/src/main/scala/org/apache/spark/scheduler/ShuffleMapTask.scala |