

Spark开发及本地环境搭建指南

www.linuxidc.com

欢迎点击这里的链接进入精彩的[Linux公社](http://www.Linuxidc.com)网站

Linux公社（www.Linuxidc.com）于2006年9月25日注册并开通网站，Linux现在已经成为一种广受关注和支持的一种操作系统，IDC是互联网数据中心，LinuxIDC就是关于Linux的数据中心。

[Linux公社](http://www.Linuxidc.com)是专业的Linux系统门户网站，实时发布最新Linux资讯，包括Linux、Ubuntu、Fedora、RedHat、红旗Linux、Linux教程、Linux认证、SUSE Linux、Android、Oracle、Hadoop、CentOS、MySQL、Apache、Nginx、Tomcat、Python、Java、C语言、OpenStack、集群等技术。

Linux公社（LinuxIDC.com）设置了有一定影响力的Linux专题栏目。

包括：[Ubuntu 专题](#) [Fedora 专题](#) [Android 专题](#) [Oracle 专题](#) [Hadoop 专题](#) [RedHat 专题](#) [SUSE 专题](#) [红旗 Linux 专题](#) [CentOS 专题](#)



目录

- 构建本机开发环境
- 向社区提交PR

欢迎点击这里的链接进入精彩的[Linux公社](http://www.Linuxidc.com)网站

Linux公社（www.Linuxidc.com）于2006年9月25日注册并开通网站，Linux现在已经成为一种广受关注和支持的一种操作系统，IDC是互联网数据中心，LinuxIDC就是关于Linux的数据中心。

[Linux公社](http://www.Linuxidc.com)是专业的Linux系统门户网站，实时发布最新Linux资讯，包括Linux、Ubuntu、Fedora、RedHat、红旗Linux、Linux教程、Linux认证、SUSE Linux、Android、Oracle、Hadoop、CentOS、MySQL、Apache、Nginx、Tomcat、Python、Java、C语言、OpenStack、集群等技术。

Linux公社（LinuxIDC.com）设置了有一定影响力的Linux专题栏目。

包括：[Ubuntu 专题](#) [Fedora 专题](#) [Android 专题](#) [Oracle 专题](#) [Hadoop 专题](#) [RedHat 专题](#) [SUSE 专题](#) [红旗 Linux 专题](#) [CentOS 专题](#)



构建本机上的Spark开发环境

www.linuxidc.com

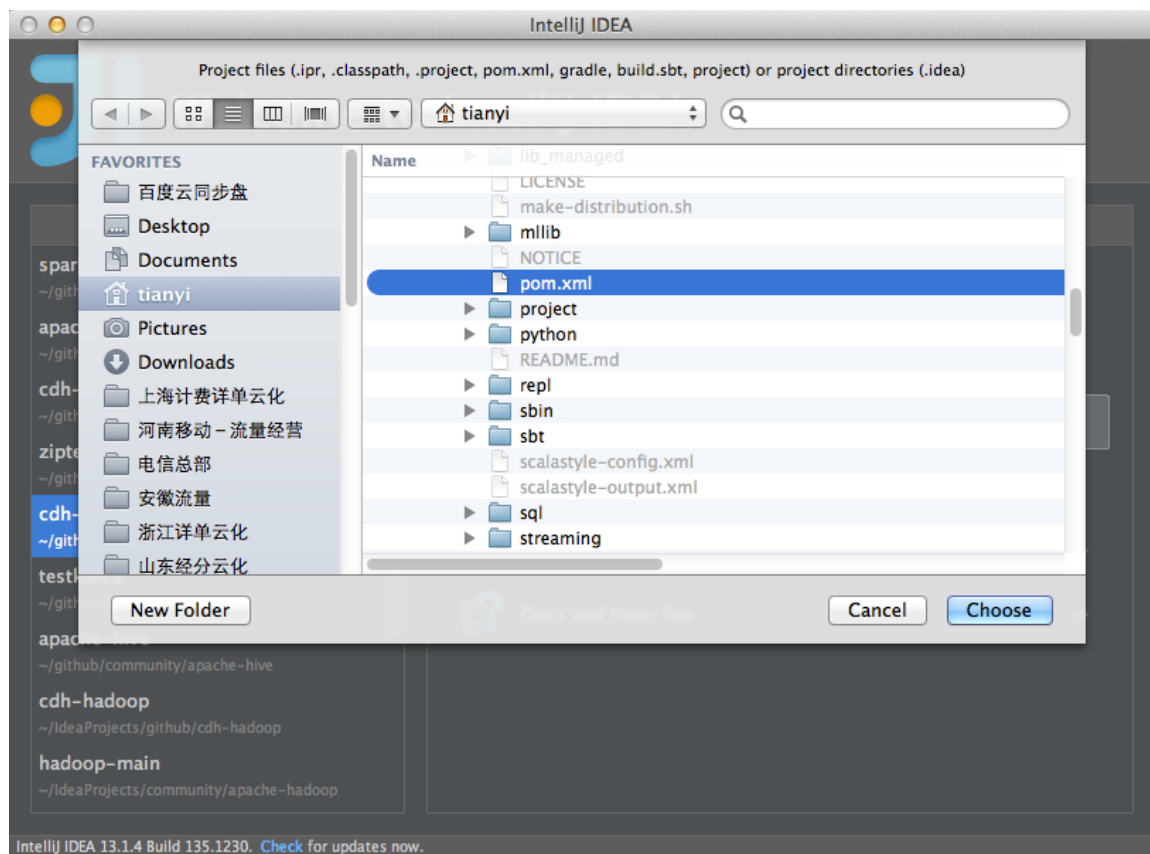
使用IDEA进行Spark开发与调试

- 环境准备
 - 推荐使用CentOS, Redhat, Fedora等Linux操作系统
 - 推荐使用MacOS
 - 如果使用PySpark (python on Spark) 需要使用JDK1.6.x
 - 安装IntelliJ IDEA (后续使用IDEA13举例)
 - 安装Scala 2.10.4
 - 安装Maven
 - 安装git客户端

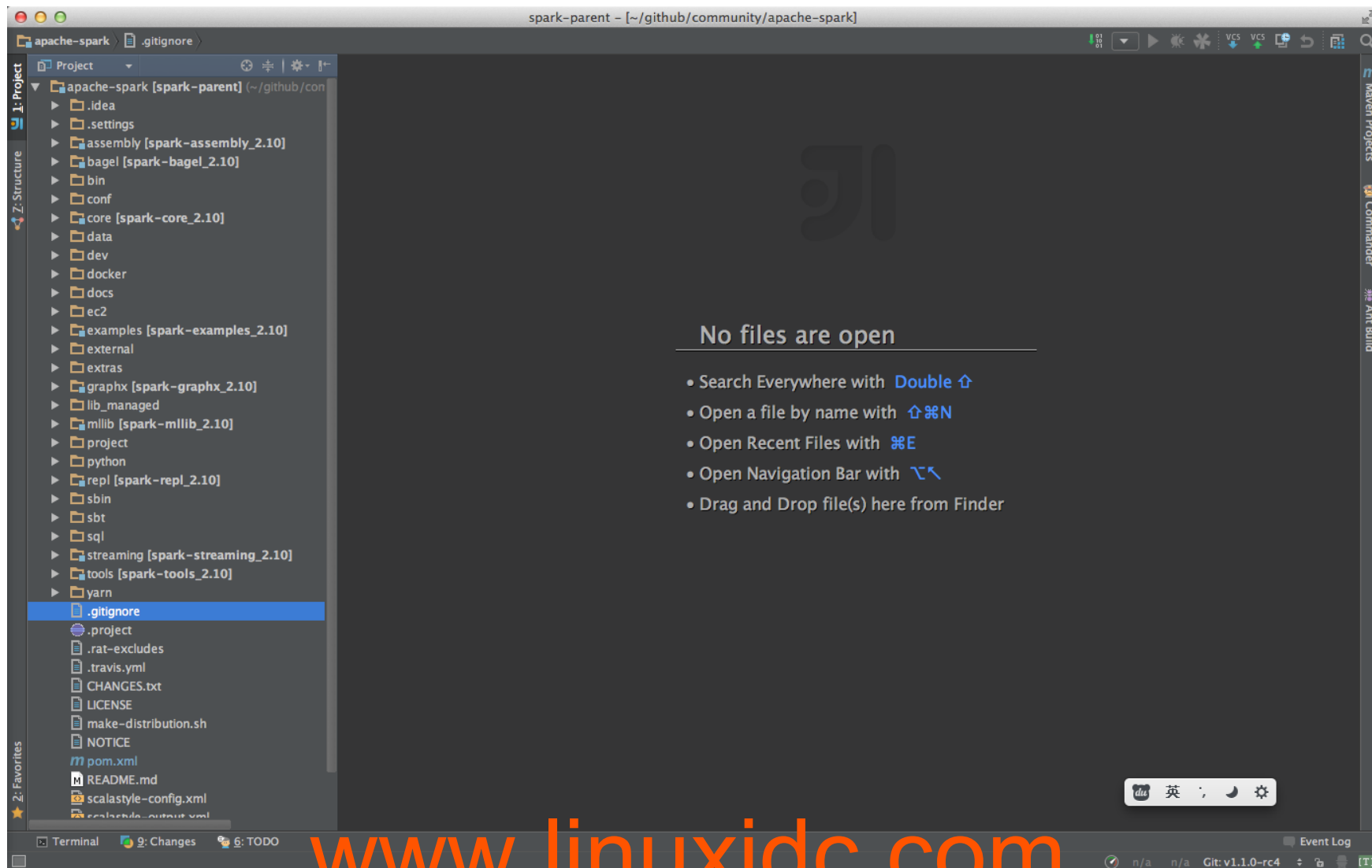
导入Spark工程

使用git命令下载源码

IDEA中选择Open Project
选择Spark目录下面的
pom.xml



导入完成



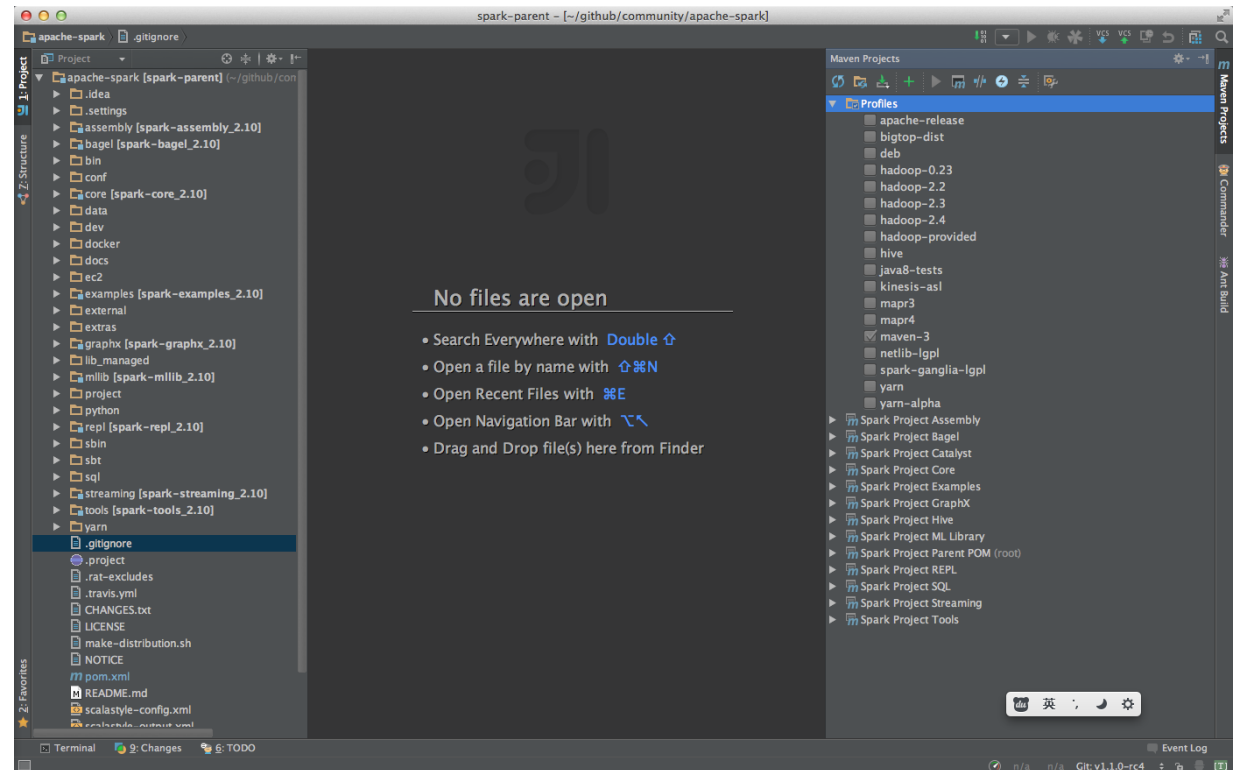
根据需求选择Profile

Hadoop版本

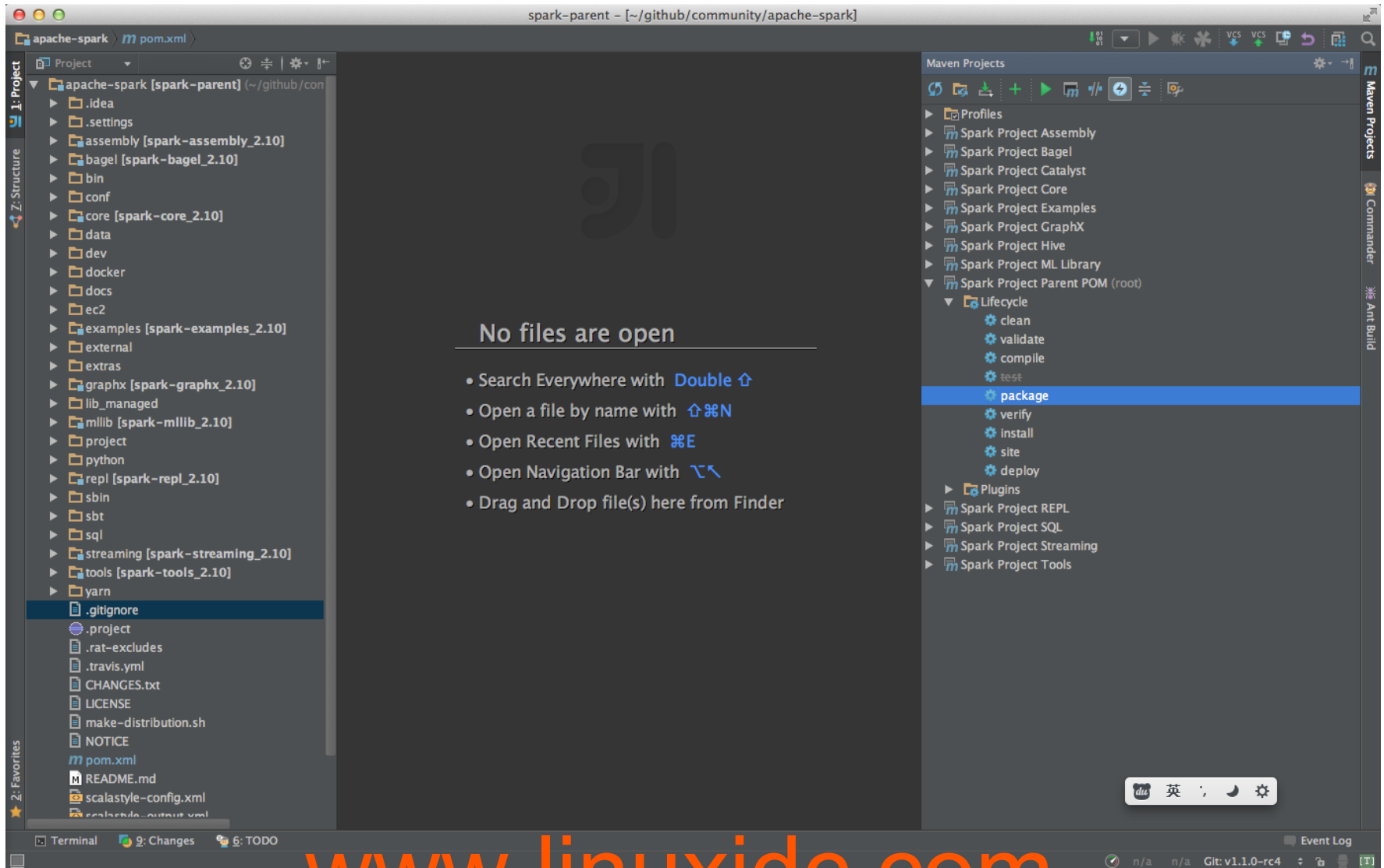
是否引入spark-hive

是否运行在Yarn

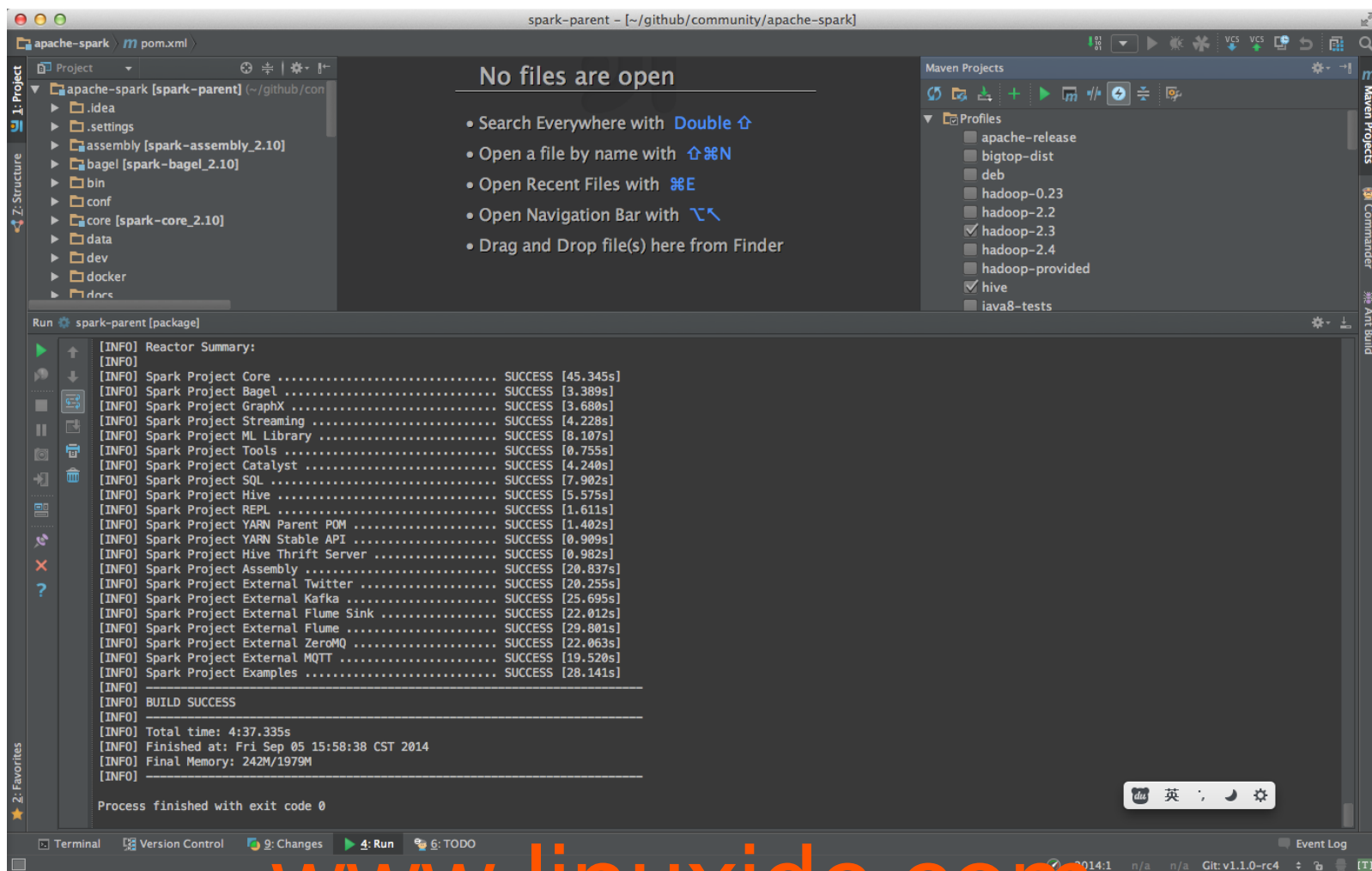
是否引入ganglia组件



使用maven编译



Maven编译成功



欢迎点击这里的链接进入精彩的[Linux公社](http://www.Linuxidc.com) 网站

Linux公社（www.Linuxidc.com）于2006年9月25日注册并开通网站，Linux现在已经成为一种广受关注和支持的一种操作系统，IDC是互联网数据中心，LinuxIDC就是关于Linux的数据中心。

[Linux公社](http://www.Linuxidc.com)是专业的Linux系统门户网站，实时发布最新Linux资讯，包括Linux、Ubuntu、Fedora、RedHat、红旗Linux、Linux教程、Linux认证、SUSE Linux、Android、Oracle、Hadoop、CentOS、MySQL、Apache、Nginx、Tomcat、Python、Java、C语言、OpenStack、集群等技术。

Linux公社（LinuxIDC.com）设置了有一定影响力的Linux专题栏目。

包括：[Ubuntu 专题](#) [Fedora 专题](#) [Android 专题](#) [Oracle 专题](#) [Hadoop 专题](#) [RedHat 专题](#) [SUSE 专题](#) [红旗 Linux 专题](#) [CentOS 专题](#)



编译准备

- 目前Spark的代码中存在一些BUG会导致IDEA无法直接编译Spark代码
- Yarn模块 (本地调试可不勾选此Profile)
- FlumeSink模块(用不到的人可以在pom.xml中删除Flume相关的module定义)

欢迎点击这里的链接进入精彩的[Linux公社](http://www.Linuxidc.com)网站

Linux公社（www.Linuxidc.com）于2006年9月25日注册并开通网站，Linux现在已经成为一种广受关注和支持的一种操作系统，IDC是互联网数据中心，LinuxIDC就是关于Linux的数据中心。

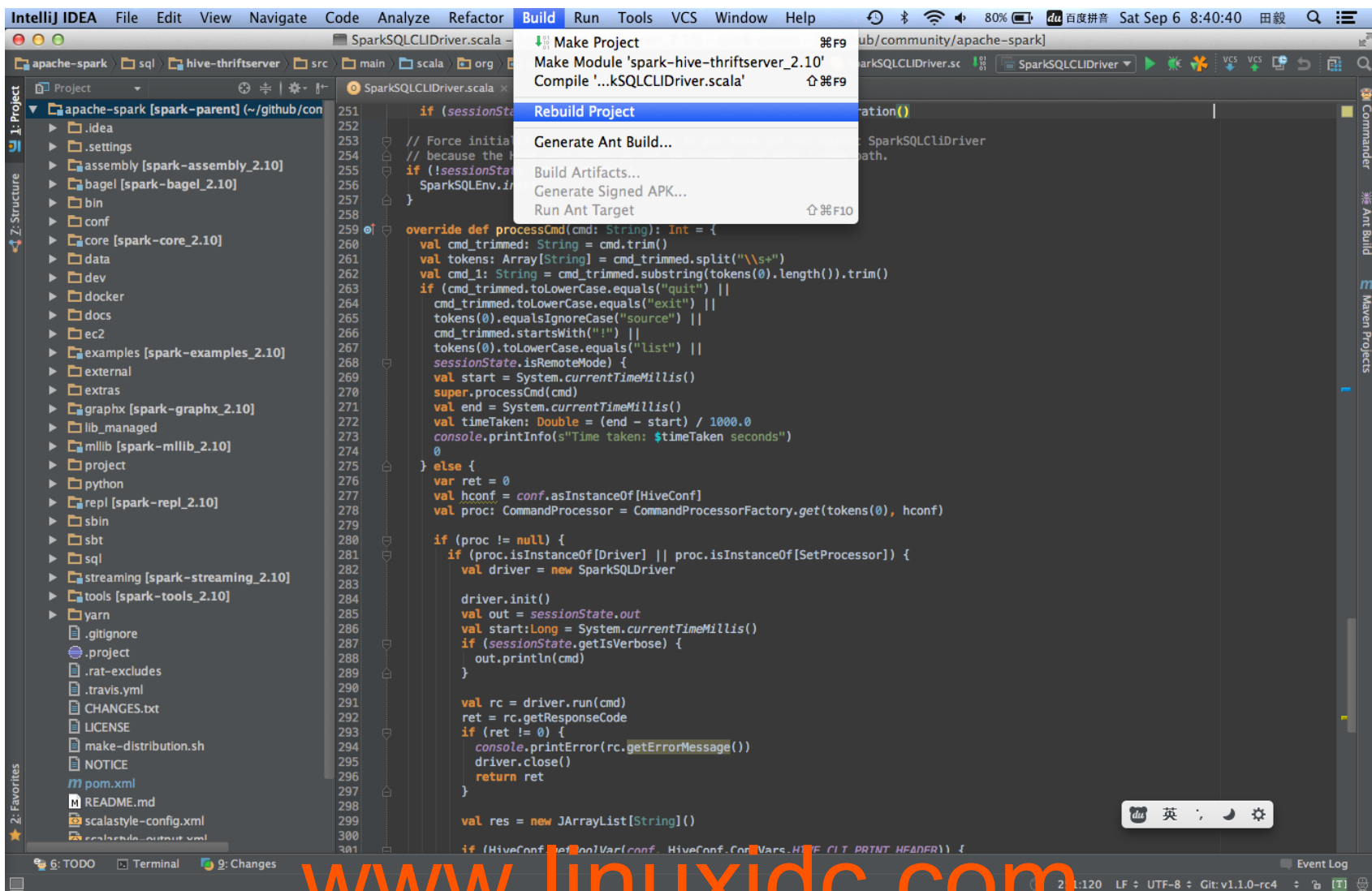
[Linux公社](http://www.Linuxidc.com)是专业的Linux系统门户网站，实时发布最新Linux资讯，包括Linux、Ubuntu、Fedora、RedHat、红旗Linux、Linux教程、Linux认证、SUSE Linux、Android、Oracle、Hadoop、CentOS、MySQL、Apache、Nginx、Tomcat、Python、Java、C语言、OpenStack、集群等技术。

Linux公社（LinuxIDC.com）设置了有一定影响力的Linux专题栏目。

包括：[Ubuntu 专题](#) [Fedora 专题](#) [Android 专题](#) [Oracle 专题](#) [Hadoop 专题](#) [RedHat 专题](#) [SUSE 专题](#) [红旗 Linux 专题](#) [CentOS 专题](#)

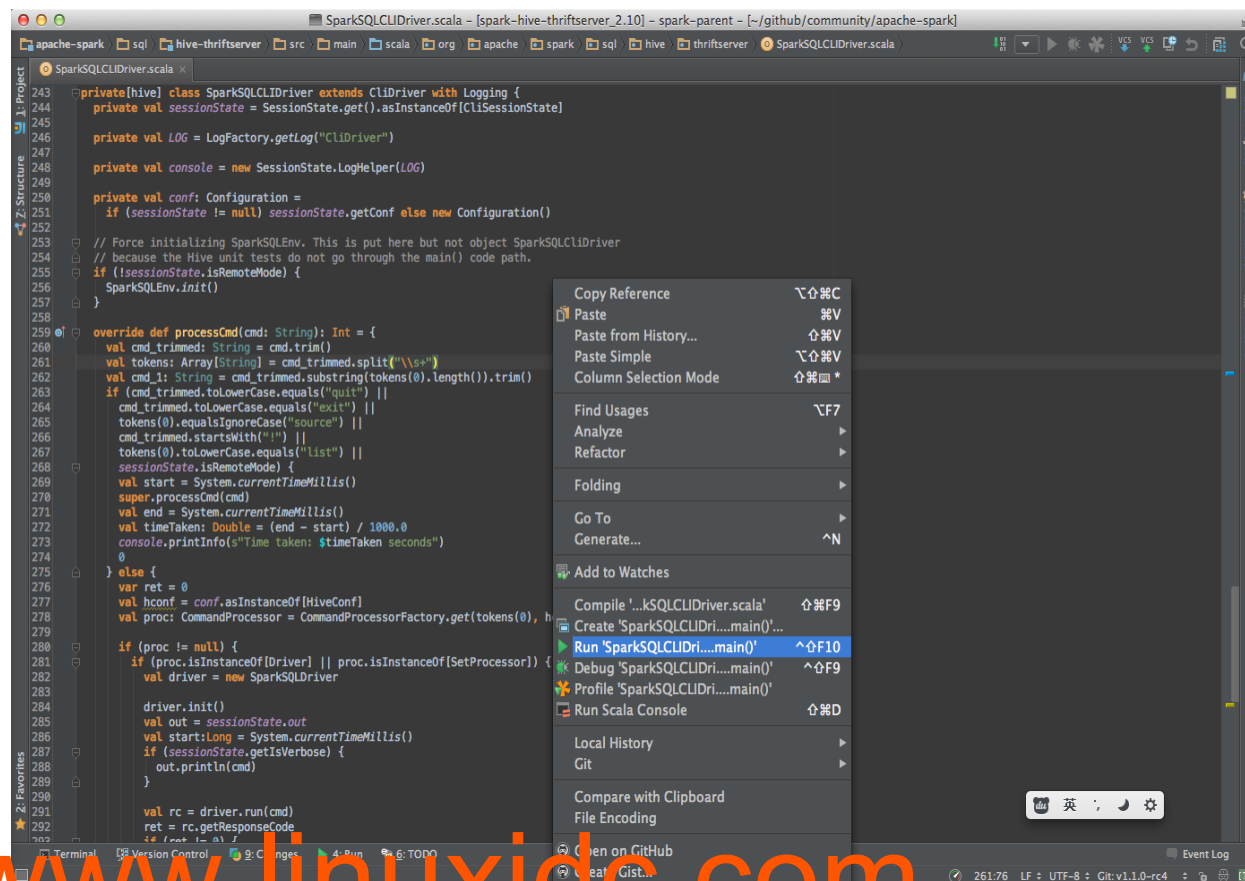


使用IDEA编译

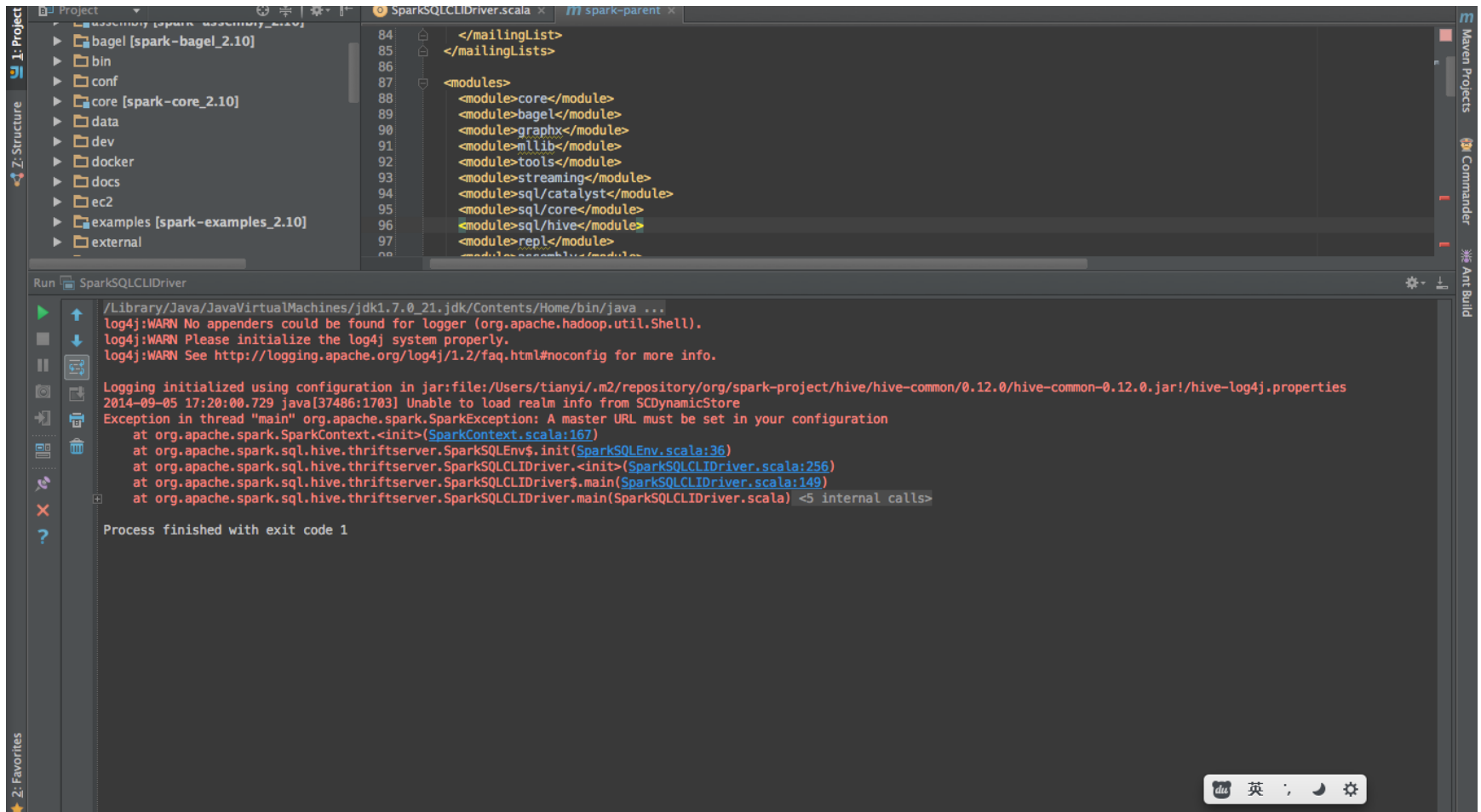


调试(以SparkSQL举例)

- 打开SparkSQLCliDriver. scala
- 右键点击 “Run ...”



第一次启动失败



The screenshot shows an IDE with a project structure on the left and a code editor in the center. The project structure includes folders like `bagel`, `bin`, `conf`, `core`, `data`, `dev`, `docker`, `docs`, `ec2`, `examples`, and `external`. The code editor displays an XML file with a `<modules>` section listing various modules like `core`, `bagel`, `graphx`, `mllib`, `tools`, `streaming`, `sql/catalyst`, `sql/core`, `sql/hive`, and `repl`. Below the code editor, a run console shows the output of a `SparkSQLCLIDriver` run. The output includes log4j warnings, logging initialization, and a fatal exception: `org.apache.spark.SparkException: A master URL must be set in your configuration`. The process finished with exit code 1.

```
Run SparkSQLCLIDriver

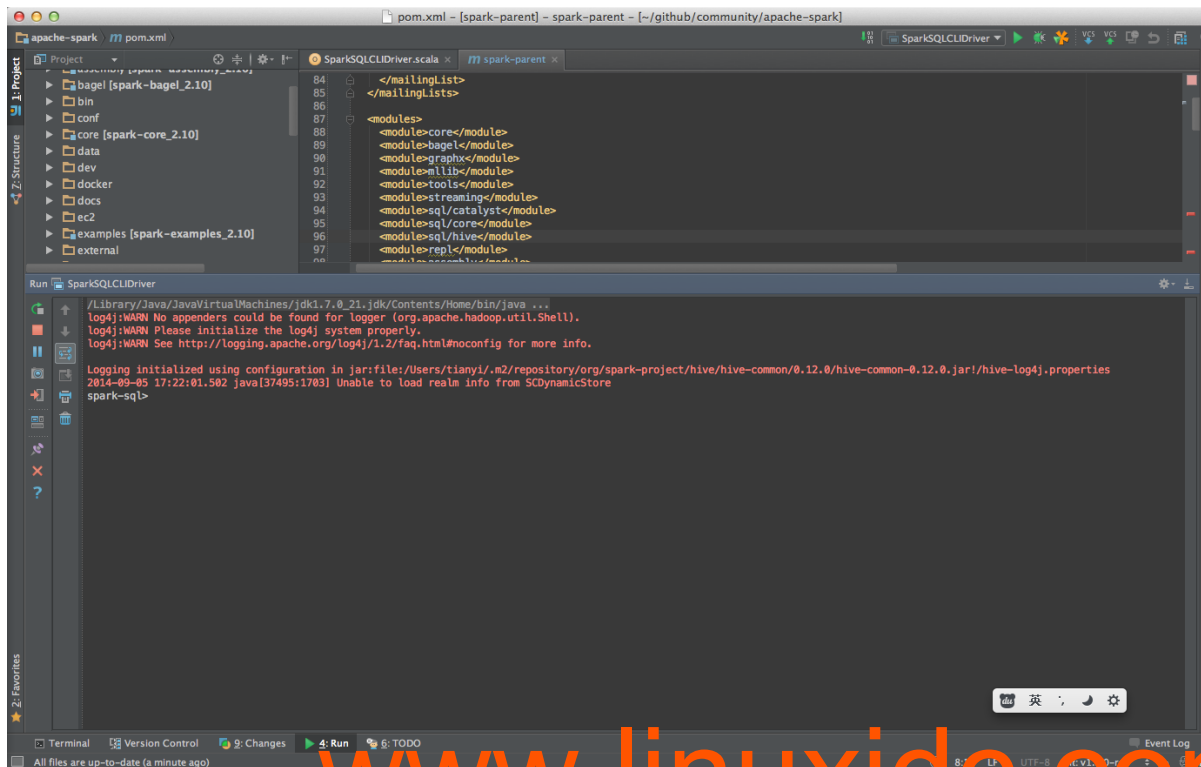
/Library/Java/JavaVirtualMachines/jdk1.7.0_21.jdk/Contents/Home/bin/java ...
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.

Logging initialized using configuration in jar:file:/Users/tianyi/.m2/repository/org/spark-project/hive/hive-common/0.12.0/hive-common-0.12.0.jar!/hive-log4j.properties
2014-09-05 17:20:00.729 java[37486:1703] Unable to load realm info from SCDynamicStore
Exception in thread "main" org.apache.spark.SparkException: A master URL must be set in your configuration
    at org.apache.spark.SparkContext.<init>(SparkContext.scala:167)
    at org.apache.spark.sql.hive.thriftserver.SparkSQLEnv$.init(SparkSQLEnv.scala:36)
    at org.apache.spark.sql.hive.thriftserver.SparkSQLCLIDriver.<init>(SparkSQLCLIDriver.scala:256)
    at org.apache.spark.sql.hive.thriftserver.SparkSQLCLIDriver$.main(SparkSQLCLIDriver.scala:149)
    at org.apache.spark.sql.hive.thriftserver.SparkSQLCLIDriver.main(SparkSQLCLIDriver.scala) <5 internal calls>

Process finished with exit code 1
```

修改启动VM参数

- 增加-Dspark.master=local[4]指定local模式
- 增加-Xmx4096m增加内存上限



DEBUG时需要注意

- 断点不要设置过多，调试Scala程序的断点开销远远大于Java，超过2个断点就会使你的程序慢的要死
- 如果需要增加Debug日志，可以将一个hive-log4j.properties文件拷贝到classpath对应的目录下面

搭建本地DEBUG环境的好处

- 快速了解程序运行的流程
- 对于解决Spark的BUG非常有用
- 本身Intellij IDEA提供了很多快捷功能, 减少敲代码编译等工作

Github上贡献自己的代码

www.linuxidc.com

创建github帐号

- 略

fork社区项目

This repository ▾ Search or type a command ⓘ Explore Gist Blog Help tianyi + - ✂ 📄

apache / spark
mirrored from [git://git.apache.org/spark.git](https://git.apache.org/spark.git)

Watch ▾ 252 ★ Star 1,060 **Fork** 1,106

Mirror of Apache Spark

7,850 commits 11 branches 24 releases 274 contributors

branch: master ▾ spark / +

[SPARK-2850] [SPARK-2626] [mlib] MLlib stats examples + small fixes ...

jkbradley authored 19 minutes ago latest commit c8b16ca0d8
→ mengxr committed 19 minutes ago

assembly	[SPARK-2410][SQL] Merging Hive Thrift/JDBC server (with Maven profile...	21 days ago
bagel	[SPARK-2410][SQL] Merging Hive Thrift/JDBC server (with Maven profile...	21 days ago
bin	[SPARK-2925] [sql]fix spark-sql and start-thriftserver shell bugs whe...	4 days ago
conf	SPARK-1902 Silence stacktrace from logs when doing port failover to p...	2 months ago
core	[SPARK-2718] [yarn] Handle quotes and other characters in user args.	4 hours ago
data	SPARK-2363. Clean MLlib's sample data files	a month ago
dev	SPARK-2884: Create binary builds in parallel with release script.	20 hours ago
docker	[SPARK-1342] Scala 2.10.4	5 months ago
docs	SPARK-3025 [SQL]: Allow JDBC clients to set a fair scheduler pool	7 hours ago

<> Code

Pull Requests 286

Pulse

Graphs

HTTPS clone URL
<https://github.com/>

You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

Clone in Desktop

Download ZIP

www.linuxidc.com

fork成功啦

The screenshot shows the GitHub interface for a repository named 'spark' by user 'tianyi'. The repository is a mirror of Apache Spark. The top navigation bar includes 'This repository', a search bar, and links to 'Explore', 'Gist', 'Blog', and 'Help'. The repository name 'tianyi / spark' is highlighted with a red box. To the right of the repository name are buttons for 'Unwatch', 'Star' (0), 'Fork' (1,106), and a share icon. Below the repository name, it says 'Mirror of Apache Spark — Edit'. A progress bar shows 7,752 commits, 11 branches, 21 releases, and 267 contributors. The main content area shows the 'master' branch, which is 99 commits behind the upstream 'apache:master'. A list of recent commits is displayed, including '[SPARK-2936] Migrate Netty network module from Java to Scala' by rxin, and '[SPARK-2410][SQL] Merging Hive Thrift/JDBC server (with Maven profile...)' by aaronrav. The right sidebar contains links for 'Code', 'Pull Requests' (0), 'Pulse', 'Graphs', and 'Settings'. At the bottom of the sidebar, there is an 'SSH clone URL' field with 'git@github.com:tianyi:spark.git' and buttons for 'Clone in Desktop' and 'Download ZIP'.

GitHub repository page for **tianyi / spark** (Mirror of Apache Spark).

Repository statistics: 7,752 commits, 11 branches, 21 releases, 267 contributors.

Current branch: **master** (spark / +). This branch is 99 commits behind apache:master.

Recent commits:

- [SPARK-2936] Migrate Netty network module from Java to Scala (latest commit ba28a8fcbc)
- [SPARK-2410][SQL] Merging Hive Thrift/JDBC server (with Maven profile...
- [SPARK-2894] spark-shell doesn't accept flags
- SPARK-1902 Silence stacktrace from logs when doing port failover to p...
- [SPARK-2936] Migrate Netty network module from Java to Scala
- SPARK-2363. Clean MLib's sample data files
- [SPARK-2894] spark-shell doesn't accept flags
- [SPARK-1342] Scala 2.10.4
- [SPARK-2635] Fix race condition at SchedulerBackend.isReady in standa...
- SPARK-2346: Add user data option to EC2 scripts

Right sidebar options: Code, Pull Requests (0), Pulse, Graphs, Settings, SSH clone URL (git@github.com:tianyi:spark.git), Clone in Desktop, Download ZIP.

用社区代码库创建本地仓库

```
TimMacBook:community tianyi$ git clone https://github.com/apache/spark test-spark
Cloning into 'test-spark'...
remote: Counting objects: 117249, done.
remote: Compressing objects: 100% (37446/37446), done.
remote: Total 117249 (delta 53990), reused 117247 (delta 53989)
Receiving objects: 100% (117249/117249), 72.56 MiB | 900.00 KiB/s, done.
Resolving deltas: 100% (53990/53990), done.
Checking connectivity... done.
Checking out files: 100% (13459/13459), done.
```

加入自己的github代码库

```
TimMacBook:test-spark tianyi$ git remote -v
origin      https://github.com/apache/spark (fetch)
origin      https://github.com/apache/spark (push)
TimMacBook:test-spark tianyi$ git remote add tianyi https://github.com/tianyi/spark
TimMacBook:test-spark tianyi$ git remote -v
origin      https://github.com/apache/spark (fetch)
origin      https://github.com/apache/spark (push)
tianyi      https://github.com/tianyi/spark (fetch)
tianyi      https://github.com/tianyi/spark (push)
```

下载自己的github代码库

```
TimMacBook:test-spark tianyi$ git fetch tianyi
remote: Counting objects: 278, done.
remote: Compressing objects: 100% (109/109), done.
remote: Total 278 (delta 67), reused 259 (delta 66)
Receiving objects: 100% (278/278), 285.16 KiB | 173.00 KiB/s, done.
Resolving deltas: 100% (67/67), done.
From https://github.com/tianyi/spark
* [new branch]    branch-0.5 -> tianyi/branch-0.5
* [new branch]    branch-0.6 -> tianyi/branch-0.6
* [new branch]    branch-0.7 -> tianyi/branch-0.7
* [new branch]    branch-0.8 -> tianyi/branch-0.8
* [new branch]    branch-0.9 -> tianyi/branch-0.9
* [new branch]    branch-1.0 -> tianyi/branch-1.0
* [new branch]    branch-1.0-jdbc -> tianyi/branch-1.0-jdbc
* [new branch]    master -> tianyi/master
* [new branch]    scala-2.9 -> tianyi/scala-2.9
* [new branch]    spark-2817 -> tianyi/spark-2817
* [new branch]    streaming -> tianyi/streaming
* [new tag]       v1.0.1-rc1 -> v1.0.1-rc1
* [new tag]       v1.0.1-rc2 -> v1.0.1-rc2
```

准备工作完成

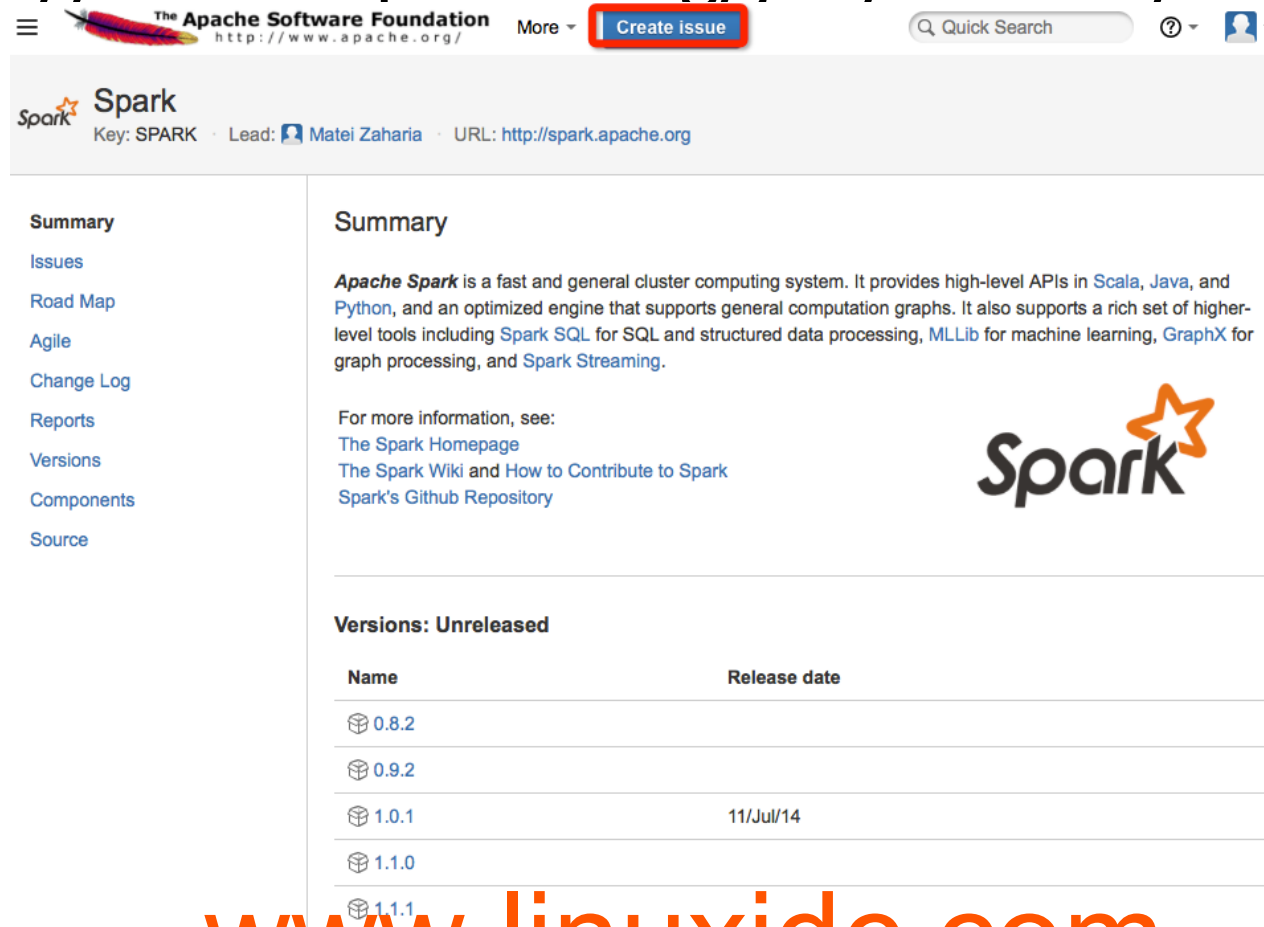
至此，你的本地开发环境已经准备好啦！

下面介绍如何提交一个Pull Request(后面简称PR)

www.linuxidc.com

在JIRA上提交一个Issue

<https://issues.apache.org/jira/browse/SPARK>



The screenshot shows the Apache Spark JIRA issue page. At the top, there's a navigation bar with the Apache Software Foundation logo, a 'Create issue' button highlighted with a red box, and a search bar. Below this is a header for 'Spark' with the key 'SPARK', lead 'Matei Zaharia', and URL 'http://spark.apache.org'. The left sidebar contains a list of links: Summary, Issues, Road Map, Agile, Change Log, Reports, Versions, Components, and Source. The main content area has a 'Summary' section with a description of Apache Spark, a link to 'The Spark Homepage', and a 'Versions: Unreleased' table. The table lists versions 0.8.2, 0.9.2, 1.0.1, 1.1.0, and 1.1.1, with release dates for 1.0.1 and 1.1.0. The Spark logo is also visible on the right side of the summary section.

Summary

Apache Spark is a fast and general cluster computing system. It provides high-level APIs in [Scala](#), [Java](#), and [Python](#), and an optimized engine that supports general computation graphs. It also supports a rich set of higher-level tools including [Spark SQL](#) for SQL and structured data processing, [MLLib](#) for machine learning, [GraphX](#) for graph processing, and [Spark Streaming](#).

For more information, see:
[The Spark Homepage](#)
[The Spark Wiki](#) and [How to Contribute to Spark](#)
[Spark's Github Repository](#)

Versions: Unreleased

Name	Release date
0.8.2	
0.9.2	
1.0.1	11/Jul/14
1.1.0	
1.1.1	

www.linuxidc.com

如何填写ISSUE表格

Summary: 总结

描述这个Issue, BUG描述或功能描述

Issue Type: 类型

- Bug
- New Feature
- Improvement

Priority: 优先级

- **Critical**, 指主要模块崩溃, 内存泄漏, 数据丢失等严重问题。
- **Major**, 主要功能有问题
- **Minor**, 次要功能有问题
- **Trivial**, 细微的问题, 比如界面易用性, 美观等无关紧要的缺陷

Component: 组件

Affects Version: 影响版本

如果是BUG, 请描述这个BUG会影响哪些已发布版本, 确认多少个版本就写多少版本

如果是New Feature, 不用写

Environment: 环境

描述BUG或新特性产生的环境条件, 如操作系统, 时区, 地区, 语言, hadoop版本等等

Description: 描述

最重要的部分, 后面专门讲

The screenshot shows a 'Create Issue' form with the following fields and options:

- Project:** Spark
- Issue Type:** Bug (with a help icon and a note: 'Some issue types are unavailable due to incompatible field configuration and/or workflow association')
- Summary:** (empty text box)
- Priority:** Major (with a help icon)
- Component/s:** (empty dropdown menu with a note: 'Start typing to get a list of possible matches or press down to select.')
- Affects Version/s:** (empty dropdown menu with a note: 'Start typing to get a list of possible matches or press down to select.')
- Fix Version/s:** (empty dropdown menu with a note: 'Start typing to get a list of possible matches or press down to select.')
- Target Version/s:** (empty dropdown menu with a note: 'Start typing to get a list of possible matches or press down to select. The versions where this patch is intended to be committed. Use "Fix Version" to note where it act committed.')
- Environment:** (empty text box)
- Footer:** ☐ Create another

如果解决的是一个BUG

- 尽量详细地描述bug的
 - 重现步骤
 - 症状(异常、stacktrace)
 - 可能的原因
 - 可能的解决方案
 - 可能相关的其他issue、PR
- 学习 <https://issues.apache.org/jira/browse/SPARK-2129>

如果解决的是一个Feature

- 尽量详细地描述Feature的
 - 修改的设计动机
 - 与其他设计的优劣势比较
 - 可能引入的新的问题
 - 当前问题域内尚未解决的问题
 - 新API的示例

创建branch

首先, 创建一个branch来专门存放你修改后的代码.

```
git checkout -b spark-2817 origin/master
```

```
git status
```

可以看到你当前的branch已经是spark-2817

下面开始你的代码修改吧, 此处略去5万字

每天将自己的工作上传到github

本地提交

```
git commit -m "fix the main problem of this  
issue"
```

提交到远程

```
git push tianyi spark-2817
```

假设你的代码开发完了

- 我们假设你现在的代码开发完了，在你创建PR之前，务必检查以下几件事情：
 - 代码Style，缩进，空格，大小写，变量名，函数名
 - 是否包含测试Case
 - 如果是SparkSQL中新增的hive功能，还需要在提交中包含golden answer的文件
 - 最重要的是，确定能够在你的本机能够通过相关的unit test，如果你不知道哪些相关，那就都跑一遍吧

假设你上述检查都结束了

务必确认自己本地的代码已经全部上传到自己的github代码库

`git status`

在github页面上创建PR

Mirror of Apache Spark — Edit

7,852 commits

12 branches

21 releases

267 contributors

Your recently pushed branches:

spark-test (1 minute ago)

Compare & pull request

branch: spark-test

spark / +



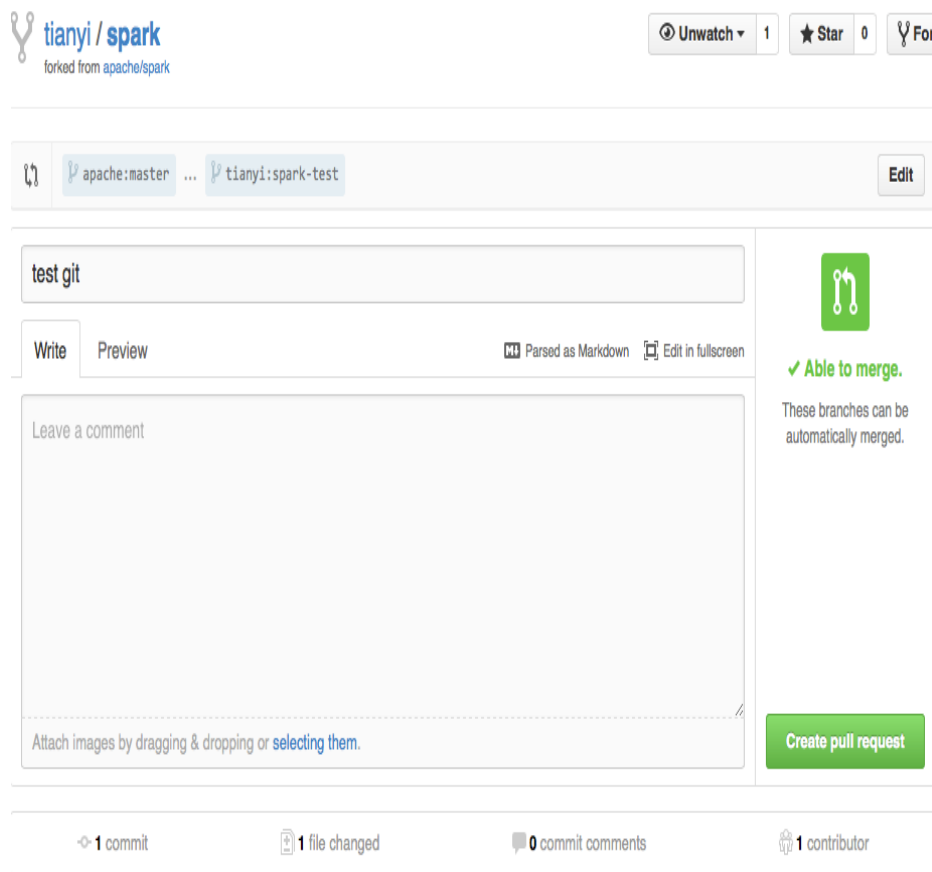
This branch is 1 commit ahead of apache:master

Pull Request Compare

test git

www.linuxidc.com

描述你的PR



title:

[ISSUE-NUM][COMPONENT] 描述

comment:

与ISSUE的描述类似, 但是不需要贴堆栈, 需要的话, 可以贴一个issue的链接进去

提交PR只是成功的一小半

- 不断的和社区的大神们交流才是王道
- 回复他们提出的各种设计，功能的疑问
- 不断修改comment中提出的代码优化建议
- 在PR追加commit只需要直接在你github代码库上commit就行，例如：

```
git commit -m "fix the code style problems"
```

```
git push tianyi spark-2817
```

沟通过程建议

- 请耐心等待, Spark社区每天有大量的PR需要review, Committer也是人, 需要一个一个来, 他们还有自己的本职工作
- 注意时差问题, 每天最好的沟通时间是早上和晚上, 注意跟踪自己的PR
- Josh提交了一个PR的Dashboard
<http://spark-prs.appspot.com> (需要翻墙)
可以很方便的跟踪社区进展

当你的PR被merge

- 恭喜你加入Spark Contributor列表
- 你的名字将出现在Spark的release列表中

一些其他问题

- 提交PR不是主要目的, 主要在于在过程中不断学习新的知识, 学习大神们的代码思路, 不断熟悉Spark的各模块功能
- 虽说人脑的纠错能力强.....但是英文还是尽量努力提升
- 日常多去看看JIRA上和Github上其他人的Issue和PR, 适当参与code review和问题解答
- 订阅user邮件列表, 能够帮助别人的同时, 提高自己

谢谢

www.linuxidc.com

欢迎点击这里的链接进入精彩的[Linux公社](http://www.Linuxidc.com)网站

Linux公社（www.Linuxidc.com）于2006年9月25日注册并开通网站，Linux现在已经成为一种广受关注和支持的一种操作系统，IDC是互联网数据中心，LinuxIDC就是关于Linux的数据中心。

[Linux公社](http://www.Linuxidc.com)是专业的Linux系统门户网站，实时发布最新Linux资讯，包括Linux、Ubuntu、Fedora、RedHat、红旗Linux、Linux教程、Linux认证、SUSE Linux、Android、Oracle、Hadoop、CentOS、MySQL、Apache、Nginx、Tomcat、Python、Java、C语言、OpenStack、集群等技术。

Linux公社（LinuxIDC.com）设置了有一定影响力的Linux专题栏目。

包括：[Ubuntu 专题](#) [Fedora 专题](#) [Android 专题](#) [Oracle 专题](#) [Hadoop 专题](#) [RedHat 专题](#) [SUSE 专题](#) [红旗 Linux 专题](#) [CentOS 专题](#)

