Spark
亚太研究院

Spark 官方文档翻译

给 Spark 提交代码

(v1.1.0)

翻译者　刘亚卿

Spark 官方文档翻译团成员

前 言

世界上第一个Spark 1.1.0 中文文档问世了！

伴随着大数据相关技术和产业的逐步成熟，继Hadoop之后，Spark技术以集大成的无可比拟的优势，发展迅速，将成为替代Hadoop的下一代云计算、大数据核心技术。

Spark是当今大数据领域最活跃最热门的高效大数据通用计算平台，基于RDD，Spark成功的构建起了一体化、多元化的大数据处理体系，在“One Stack to rule them all”思想的引领下，Spark成功的使用Spark SQL、Spark Streaming、MLLib、GraphX近乎完美的解决了大数据中Batch Processing、Streaming Processing、Ad-hoc Query等三大核心问题，更为美妙的是在Spark中Spark SQL、Spark Streaming、MLLib、GraphX四大子框架和库之间可以无缝的共享数据和操作，这是当今任何大数据平台都无可匹敌的优势。

在实际的生产环境中，世界上已经出现很多一千个以上节点的Spark集群，以eBay为例，eBay的Spark集群节点已经超过2000个，Yahoo 等公司也在大规模的使用Spark，国内的淘宝、腾讯、百度、网易、京东、华为、大众点评、优酷土豆等也在生产环境下深度使用Spark。2014 Spark Summit上的信息，Spark已经获得世界 20 家顶级公司的支持，这些公司中包括Intel、IBM等，同时更重要的是包括了最大的四个Hadoop发行商，都提供了对Spark非常强有力的支持。

与Spark火爆程度形成鲜明对比的是Spark人才的严重稀缺，这一情况在中国尤其严重，这种人才的稀缺，一方面是由于Spark技术在 2013、2014 年才在国内的一些大型企业里面被逐步应用，另一方面是由于匮乏Spark相关的中文资料和系统化的培训。为此，Spark亚太研究院和 51CTO联合推出了“Spark亚太研究院决胜大数据时代 100 期公益大讲堂”，来推动Spark技术在国内的普及及落地。

具体视频信息请参考 http://edu.51cto.com/course/course_id-1659.html

与此同时，为了向Spark学习者提供更为丰富的学习资料，Spark亚太研究院发起并号召，结合网络社区的力量构建了Spark中文文档专家翻译团队，历经 1 个月左右的艰苦努力和反复修改，Spark中文文档V1.1 终于完成。尤其值得一提的是，在此次中文文档的翻译期间，Spark官方团队发布了Spark 1.1.0 版本，为了让学习者了解到最新的内容，Spark中文文档专家翻译团队主动提出基于最新的Spark 1.1.0 版本，更新了所有已完成的翻译内容，在此，我谨代表Spark亚太研究院及广大Spark学习爱好者向专家

翻译团队所有成员热情而专业的工作致以深刻的敬意！

当然，作为世界上第一份相对系统的Spark中文文档，不足之处在所难免，大家有任何建议或者意见都可以发邮件到marketing@sparkinchina.com ;同时如果您想加入Spark中文文档翻译团队，也请发邮件到marketing@sparkinchina.com进行申请；Spark中文文档的翻译是一个持续更新的、不断版本迭代的过程，我们会尽全力给大家提供更高质量的Spark中文文档翻译。

最后，也是最重要的，请允许我荣幸的介绍一下我们的Spark中文文档第一个版本翻译的专家团队成员，他们分别是（排名不分先后）:

- 傅智勇, 《快速开始(v1.1.0)》（和唐海东翻译的是同一主题，大家可以对比参考）
- 吴洪泽,《Spark机器学习库 (v1.1.0)》（其中聚类和降维部分是蔡立宇翻译）
- 武扬，《在Yarn上运行Spark (v1.1.0)》《Spark 调优(v1.1.0)》
- 徐骄，《Spark配置(v1.1.0)》《Spark SQL编程指南(v1.1.0)》（Spark SQL和韩保礼翻译的是同一主题，大家可以对比参考）
- 蔡立宇，《Bagel 编程指南(v1.1.0)》
- harli，《Spark 编程指南 (v1.1.0)》
- 吴卓华，《图计算编程指南(1.1.0)》
- 樊登贵，《EC2(v1.1.0)》《Mesos(v1.1.0)》
- 韩保礼，《Spark SQL编程指南(v1.1.0)》（和徐骄翻译的是同一主题，大家可以对比参考）
- 颜军，《文档首页(v1.1.0)》
- Jack Niu，《Spark实时流处理编程指南(v1.1.0)》
- 俞杭军，《sbt-assembly》《使用Maven编译Spark(v1.1.0)》
- 唐海东，《快速开始(v1.1.0)》（和傅智勇翻译的是同一主题，大家可以对比参考）
- 刘亚卿，《硬件配置(v1.1.0)》《Hadoop 第三方发行版(v1.1.0)》《给Spark提交代码(v1.1.0)》
- 耿元振《集群模式概览(v1.1.0)》《监控与相关工具(v1.1.0)》《提交应用程序(v1.1.0)》
- 王庆刚，《Spark作业调度(v1.1.0)》《Spark安全(v1.1.0)》
- 徐敬丽，《Spark Standalone 模式 (v1.1.0)》

另外关于Spark API的翻译正在进行中，敬请大家关注。

Life is short, You need Spark!

Spark亚太研究院院长 王家林
2014 年 10 月

Spark 亚太研究院决胜大数据时代 100 期公益大讲堂

简　介

作为下一代云计算的核心技术，Spark性能超Hadoop百倍，算法实现仅有其 1/10 或 1/100,是可以革命Hadoop的目前唯一替代者，能够做Hadoop做的一切事情，同时速度比Hadoop快了 100 倍以上。目前Spark已经构建了自己的整个大数据处理生态系统，国外一些大型互联网公司已经部署了Spark。甚至连Hadoop的早期主要贡献者Yahoo现在也在多个项目中部署使用Spark；国内的淘宝、优酷土豆、网易、Baidu、腾讯、皮皮网等已经使用Spark技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark正在逐渐走向成熟，并在这个领域扮演更加重要的角色，刚刚结束的 2014 Spark Summit上的信息，Spark已经获得世界 20 家顶级公司的支持，这些公司中包括Intel、IBM等，同时更重要的是包括了最大的四个Hadoop发行商都提供了对非常强有力的支持Spark的支持.

鉴于Spark的巨大价值和潜力，同时由于国内极度缺乏Spark人才，Spark亚太研究院在完成了对Spark源码的彻底研究的同时，不断在实际环境中使用Spark的各种特性的基础之上，推出了Spark亚太研究院决胜大数据时代 100 期公益大讲堂，希望能够帮助大家了解Spark的技术。同时，对Spark人才培养有近一步需求的企业和个人，我们将以公开课和企业内训的方式，来帮助大家进行Spark技能的提升。同样，我们也为企业提供一体化的顾问式服务及Spark一站式项目解决方案和实施方案。

Spark亚太研究院决胜大数据时代 100 期公益大讲堂是国内第一个Spark课程免费线上讲座，每周一期，从 7 月份起，每周四晚 20:00-21:30，与大家不见不散！老师将就Spark内核剖析、源码解读、性能优化及商业实战案例等精彩内容与大家分享，干货不容错过！

时间：从 7 月份起，每周一期，每周四晚 20:00-21:30
形式：腾讯课堂在线直播
学习条件：对云计算大数据感兴趣的技术人员
课程学习地址：http://edu.51cto.com/course/course_id-1659.html

给 Spark 提交代码

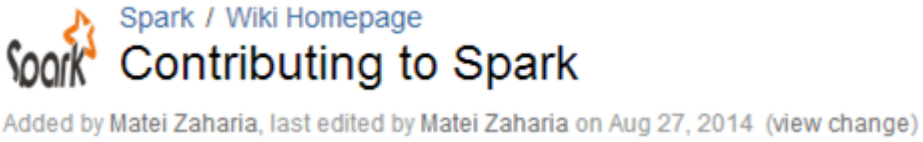
(v1.1.0)

（翻译者：刘亚卿）

Contributing to Spark，原文档链接：

https://cwiki.apache.org/confluence/display/SPARK/Contributing+to+Spark

目录



https://cwiki.apache.org/confluence/display/SPARK/Contributing+to+Spark

Apache Spark 团队欢迎大家提交各种各样的代码，不管是 bug 报告，文档，还是新的补丁。

- 问题报告
- 发布代码
- 启动任务
- 文档
- 发展探讨
- IDE 设置

1. 问题报告

如果你想报告一个 Spark 的bug或者需要一个新的特性， 在 *Apache Spark JIRA*上贴出你的问题。为了一般的用处帮助，你应该发邮件给 *user mailing list*.

2. 发布代码

我们更愿意接受以*GitHub Pull Requests*的方式发布的代码。首先在 *Spark Project JIRA*上发布一个你要更改的议题(并搞清楚这个议题是不是一个已经存在了),然后在你完成开发工作后,发送一个 Pull Requests 到 *github.com/apache/spark*存储库。

请遵循以下步骤发布你的代码：

1.把你的东西 尽可能的切分成一个个小的，单一功能的小块。合并一个带有很大改变的许多的分散的特性是很困难的。

2. 在*Spark Project JIRA* 为你的补丁建立一个议题。

3. 如果你提议了一个很大的改动 ,请附上一个设计文档并且发邮件给 dev mailing list。

4. 提交补丁作为一个GitHub pull request。请参见 forking a repo和sending a pull request. 上的GitHub 指导，那里有教程。用 JIRA名字来命名你的pull request，要包含相关的Spark 模块 或者WIP。

5. 遵循 Spark Code Style Guide 。在提交你的 pull requests 前， 你可以运行 *./dev/lint-scala* 和 *./dev/lint-python* *来验证。*

6. 确保您的代码通过自动化测试(参见下面的自动化测试)

7.添加新的测试代码。我们使用 ScalaTest 进行测试。只是在*core/src/test* 添加一个新的测试套件,或在已有的套件添加新的函数。

8.如果你添加一个新特性或配置参数，要更新这个文档(在文档文件夹)。

如果你想报告一个错误但没有时间去修复它,你仍然可以把它贴到我们的问题跟踪器上,或发邮件给我们。

Tip: Use descriptive names in your pull requests SPARK-123: Add some feature to Spark [STREAMING] SPARK-123: Add some feature to Spark streaming [MLLIB] [WIP] SPARK-123: Some potentially useful feature for MLLib

2.1 为 MLLib 贡献新的算法

虽然包含丰富的算法是 MLLib 一个重要目标,但是项目把可维护性、一致性和代码质量看成首要的。新算法应该是：

- 广为所知的。
- 已经被用的或被接受的（学术引用和具体应用案例可以帮助证明）。
- 可高度伸缩的。
- 有很好的文档支持
- 要有一些 API 跟 MLLib 中其他算法相协作。
- 要有开发者支持的合理期望。

2.2 自动测试

Spark附带了一个比较全面测试套件提供给单元测试、功能测试和综合测试。所有pull requests都在*Jenkins*自动测试,目前是由伯克利AMPLab主办的。为了能贯穿整个测试（代

码风格检查和二进制兼容性检查），运行 “*/dev/run-tests*”。

3. 启动任务

如果你是一个Spark新手并且想上传代码的话，请浏览 *list of starter tasks on our JIRA* 。这里的任务都很小很简单，但又是可以提升自己的优秀问题。

4. 文档

如果你要发布文档的话，可以通过两种方式：

- 发布一个外部链接给我们，仅邮寄给*developer mailing list*。
- 修改*built-in documentation*， 编辑在Spark'的docs文件夹中的 MarkDown 源文件，发送一个补丁到*Spark GitHub repository*。 README 文件 教你如何构建本地文档去验证你的变化。

5. 发展探讨

保持讨论的更新，加入*developer mailing list*。

6. IDE 设置

6.1 IntelliJ

尽管好多的Spark开发者会使用命令行下的 SBT 或Maven，我们使用最多的集成开发工具是 *IntelliJ IDEA*。 你可以获得免费的社区版（Apache的委托者可以获得免费的最终版的授权 *free IntelliJ Ultimate Edition licenses* ）。 从 Preferences > Plugins 安装 JetBrains Scala 插件。用IntelliJ创建一个Spark项目,只需查看存储库和使用IntelliJ的导入功能函数把Spark项目当作一个Maven项目导入到Spark中。

6.2 Eclipse

我们可以用 Eclipse 来开发和测试 Spark ，做好以下配置：

- Eclipse Juno

1 **IntelliJ**

尽管好多的Spark开发者会使用命令行下的 SBT 或Maven，我们使用最多的集成开发工具是 *IntelliJ IDEA*。 你可以获得免费的社区版（Apache的委托者可以获得免费的最终版的授权 *free IntelliJ Ultimate Edition licenses* ）。 从 Preferences > Plugins 安装 JetBrains Scala 插件。用IntelliJ创建一个Spark项目,只需查看存储库和使用IntelliJ的导入功能函数把Spark项目当作一个Maven项目导入到Spark中。

2 Eclipse

我们可以用 Eclipse 来开发和测试 Spark ，做好以下配置：

- Eclipse Juno
- Scala IDE v 3.0.3
- Scala Test

Scala IDE 在 Eclipse 的 Help > Marketplace .... search for Scala IDE 安装。记得 把 Scala Test 作为一个 Scala IDE 插件。安装完 Scala IDE 后安装 Scala Test ，按照以下步骤：

- Select Help | Install New Software
- Select http://download.scala-ide.org... in the "Work with" combo box
- Expand Scala IDE plugins, select ScalaTest for Scala IDE and install

SBT 可以创建 Eclipse .project 和 .classpath 文件. 为每一个 Spark 子项目创建文件, 用下面命令：

```
sbt/sbt eclipse
```

要导入一个特定的项目，例如 spark-core, 选择 “*File | Import | Existing Projects into Workspace*” 。不要选择 “*Copy projects into workspace*” 。不建议一下子导入所有的 Spark 子项目。ScalaTest 可以通过右击一个源文件，选择 “*Run As | Scala Test*” 执行单元测试。

如果 Java 出现内存错误，就有必要在 Eclipse 安装文件夹中的 eclipse.ini 增加设定的内存大小， 按需求增加：

```
--launcher.XXMaxPermSize
256M
```

ScalaTest 问题

如果在运行 ScalaTest 时发生如下错误：

```
An internal error occurred during:"Launching XYZSuite.scala".
java.lang.NullPointerException
```

这是因为在环境变量中设错了一个 Scala 库。右击项目，选择“*Build Path | Configure Build Path*”：

- Add Library | Scala Library
- Remove scala-library-2.10.4.jar - lib_managed\jars

出现"*Could not find resource path for Web UI: org/apache/spark/ui/static*"，是因为环境变量问题(有些类没有编译好). 要改正， 在命令行下运行:

```
sbt/sbt "test-only org.apache.spark.rdd.SortingSuite"
```

Spark 亚太研究院

Spark 亚太研究院是中国最专业的一站式大数据 Spark 解决方案供应商和高品质大数据企业级完整培训与服务供应商，以帮助企业规划、架构、部署、开发、培训和使用 Spark 为核心，同时提供 Spark 源码研究和应用技术训练。针对具体 Spark 项目，提供完整而彻底的解决方案。包括 Spark 一站式项目解决方案、Spark 一站式项目实施方案及 Spark 一体化顾问服务。

官网： www.sparkinchina.com

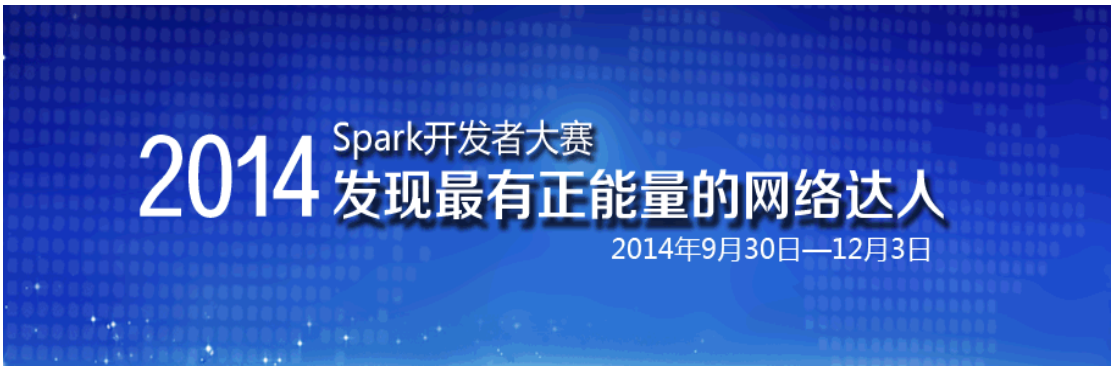
近期活动

- 2014 年亚太地区规格最高的 Spark 技术盛会！
- 面向大数据、云计算开发者、技术爱好者的饕餮盛宴！
- 云集国内外 Spark 技术领军人物及灵魂人物！
- 技术交流、应用分享、源码研究、商业案例探询！

时间：2014 年 12 月 6-7 日

地点：北京珠三角万豪酒店

Spark亚太峰会网址：http://www.sparkinchina.com/meeting/2014yt/default.asp



- 如果你是对 Spark 有浓厚兴趣的初学者，在这里你会有绝佳的入门和实践机会！
- 如果你是 Spark 的应用高手，在这里以“武”会友，和技术大牛们尽情切磋！
- 如果你是对 Spark 有深入独特见解的专家，在这里可以尽情展现你的才华！

比赛时间：

2014 年 9 月 30 日—12 月 3 日

Spark开发者大赛网址：http://www.sparkinchina.com/meeting/2014yt/dhhd.asp

视频课程：

《大数据 Spark 实战高手之路》 国内第一个 Spark 视频系列课程

从零起步，分阶段无任何障碍逐步掌握大数据统一计算平台 Spark，从 Spark 框架编写和开发语言 Scala 开始，到 Spark 企业级开发，再到 Spark 框架源码解析、Spark 与 Hadoop 的融合、商业案例和企业面试，一次性彻底掌握 Spark，成为云计算大数据时代的幸运儿和弄潮儿，笑傲大数据职场和人生！

- 第一阶段：熟练的掌握 Scala 语言
  课程学习地址：http://edu.51cto.com/pack/view/id-124.html
- 第二阶段：精通 Spark 平台本身提供给开发者 API
  课程学习地址：http://edu.51cto.com/pack/view/id-146.html
- 第三阶段：精通 Spark 内核
  课程学习地址：http://edu.51cto.com/pack/view/id-148.html
- 第四阶段：掌握基于 Spark 上的核心框架的使用
  课程学习地址：http://edu.51cto.com/pack/view/id-149.html
- 第五阶段：商业级别大数据中心黄金组合：Hadoop+ Spark
  课程学习地址：http://edu.51cto.com/pack/view/id-150.html
- 第六阶段：Spark 源码完整解析和系统定制
  课程学习地址：http://edu.51cto.com/pack/view/id-151.html

近期公开课：

《决胜大数据时代：Hadoop、Yarn、Spark 企业级最佳实践》

集大数据领域最核心三大技术：Hadoop 方向 50%：掌握生产环境下、源码级别下的 Hadoop 经验，解决性能、集群难点问题；Yarn 方向 20%：掌握最佳的分布式集群资源管理框架，能够轻松使用 Yarn 管理 Hadoop、Spark 等；Spark 方向 30%：未来统一的大数据框架平台，剖析 Spark 架构、内核等核心技术，对未来转向 SPARK 技术，做好技术储备。课程内容落地性强，即解决当下问题，又有助于驾驭未来。

开课时间： 2014 年 10 月 26-28 日北京、2014 年 11 月 1-3 日深圳

咨询电话：4006-998-758

QQ 交流群： 1 群：317540673 （已满）
2 群：297931500

微信公众号：spark-china