



Spark 官方文档翻译

使用 Maven 编译 Spark (v1.1.0)

翻译者 俞杭军

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，敬请关注。

Life is short, You need Spark!

Spark亚太研究院院长 王家林

2014 年 10 月

3 / 12

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,希望大家能够了解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

使用 Maven 编译 Spark

(v1.1.0)

(翻译者：俞杭军)

Building Spark with Maven， 原文档链接：

<http://spark.apache.org/docs/latest/building-with-maven.html>

目录

1. 使用Maven编译Spark	6
2. 设置Maven的内存使用情况	6
3. 指定Hadoop版本	7
4. 带Hive和JDBC支持的编译	8
5. 使用Maven测试Spark	8
6. 持续编译	9
7. 使用IntelliJ IDEA	9
8. 编译Spark Debian包	9
9. 运行Java 8 的测试组件	9
10. 基于Yarn编译PySpark (Building for PySpark on YARN)	10
11. 不使用Yarn的Hadoop依赖库打包	10

1. 使用 Maven 编译 Spark

- 设置 Maven 的内存使用情况
- 指定 Hadoop 版本
- 带 Hive 和 JDBC 支持的编译
- 使用 Maven 测试 Spark
- 持续编译
- 使用 IntelliJ IDEA
- 编译 Spark Debian 包
- 运行 Java 8 测试组件
- 基于 Yan 编译 PySpark
- 不使用 Hadoop 依赖库打包 Yarn

使用 Maven 编译 Spark 需要 Maven 3.0.4 或更高版本以及 Java 6 以上版本。

2. 设置 Maven 的内存使用情况

通过设置 MAVEN_OPTS 参数，需要分配比通常更多的内存来设置 Maven。我们推荐使用以下设置：

```
export MAVEN_OPTS="-Xmx2g -XX:MaxPermSize=512M  
-XX:ReservedCodeCacheSize=512m"
```

如果不运行上述命令，可能会碰到类似下面的错误：

```
[INFO] Compiling 203 Scala sources and 9 Java sources to  
/Users/me/Development/spark/core/target/scala-2.10/classes...  
[ERROR] PermGen space -> [Help 1]  
  
[INFO] Compiling 203 Scala sources and 9 Java sources to  
/Users/me/Development/spark/core/target/scala-2.10/classes...  
[ERROR] Java heap space -> [Help 1]
```

可以通过之前提到的设置 MAVEN_OPTS 变量解决这个问题。

注：对于 Java 8，上述步骤不是必需的。

3. 指定 Hadoop 版本

由于 HDFS 各版本之间协议不兼容，如果想从 HDFS 读取（文件），可以在环境变量中指定 HDFS 版本编译 Spark，即通过“hadoop.version”属性进行设置。如果未设置该属性，Spark 将默认使用 Hadoop1.0.4 版本编译。注意：某些版本的配置文件需要对应特定的 Hadoop 版本：

Hadoop 版本	所需配置文件
0.23.x	Hadoop-0.23
1.x 到 2.1.x 的	（无）
2.2.x 的	Hadoop-2.2
2.3.x	Hadoop-2.3
2.4.x 的	Hadoop-2.4

对于 Apache 的 Hadoop1.x 版本、Cloudera CDH “mr1” 发行版本、以及其他不基于 Yarn 的 Hadoop 版本，使用：

```
# Apache Hadoop 1.2.1
mvn -Dhadoop.version=1.2.1 -DskipTests clean package

# Cloudera CDH 4.2.0 with MapReduce v1
mvn -Dhadoop.version=2.0.0-mr1-cdh4.2.0 -DskipTests clean package

# Apache Hadoop 0.23.x
mvn -Phadoop-0.23 -Dhadoop.version=0.23.7 -DskipTests clean package
```

对于 Apache Hadoop 的 2.X ,0.23.x ,Cloudera CDH, 以及其他基于 Yarn 的 Hadoop 版本，可以启用“yarn-alpha” 或者“yarn” 配置文件，如果与“hadoop.version” 参数值不一致的话则可选配置“yarn.version” 属性。所需的额外配置文件取决于 Yarn 版本：

Yarn 版本	所需配置文件
0.23.x-2.1.x 版本	yarn-alpha
2.2.x 及更高版本	yarn

例如：

```
# Apache Hadoop 2.0.5-alpha
mvn -Pyarn-alpha -Dhadoop.version=2.0.5-alpha -DskipTests clean package

# Cloudera CDH 4.2.0
mvn -Pyarn-alpha -Dhadoop.version=2.0.0-cdh4.2.0 -DskipTests clean package

# Apache Hadoop 0.23.x
mvn -Pyarn-alpha -Phadoop-0.23 -Dhadoop.version=0.23.7 -DskipTests clean
package

# Apache Hadoop 2.2.X
mvn -Pyarn -Phadoop-2.2 -Dhadoop.version=2.2.0 -DskipTests clean package

# Apache Hadoop 2.3.X
mvn -Pyarn -Phadoop-2.3 -Dhadoop.version=2.3.0 -DskipTests clean package

# Apache Hadoop 2.4.X
mvn -Pyarn -Phadoop-2.4 -Dhadoop.version=2.4.0 -DskipTests clean package

# Different versions of HDFS and YARN.
mvn -Pyarn-alpha -Phadoop-2.3 -Dhadoop.version=2.3.0 -Dyarn.version=0.23.7
-DskipTests clean package
```

4. 带 Hive 和 JDBC 支持的编译

如果开启带 Hive 整合以及 JDBC 服务器和命令行界面(CLI)支持的 Spark SQL，添加 -Phive 配置参数到现有的编译选项中。

5. 使用 Maven 测试 Spark

默认使用 [ScalaTest Maven插件](#) 运行测试。

有些测试需要先将 Spark 打包，所以首次测试时运行 mvn package -DskipTests 命令。

以下是正确的执行顺序：


```
mvn -Pyarn -Phadoop-2.3 -DskipTests -Phive clean package
mvn -Pyarn -Phadoop-2.3 -Phive test
```

ScalaTest 插件还支持只运行特定的测试组件，如下：

```
mvn -Dhadoop.version=... -DwildcardSuites=org.apache.spark.repl.ReplSuite test
```

6. 持续编译

我们使用 scala-maven-plugin 插件支持渐进和持续地编译。例如：

```
mvn scala:cc
```

将进行持续编译（例如随时监测代码变化，一有改变就编译(wait for changes)）。然而，这种方式还没有广泛测试过。

7. 使用 IntelliJ IDEA

上文的安装步骤在 IntelliJ IDEA 11.1.4 上没有问题。通过项目根目录下的 pom.xml 文件打开工程，无论是 hadoop1 还是 hadoop2 配置文件，只要激活 Maven Properties 即可使用 IDEA。不过还没在 Eclipse / Scala 的 IDE 测试过。

8. 编译 Spark Debian 包

使用 Maven 编译包括支持包含 assembly ‘fat-jar’、PySpark 以及必要的脚本和配置文件的 Debian 软件包。可以通过指定下列命令进行创建包：

```
mvn -Pdeb -DskipTests clean package
```

可以在执行 assembly/target 命令后得到 Debian 软件包。我们将 short commit hash 添加到文件名中，从而可以区分出单独的包用于编译 SNAPSHOT 版本。

9. 运行 Java 8 的测试组件

仅支持 Java 8

```
mvn install -DskipTests -Pjava8-tests
```

仅当 `-Pjava8-tests` 配置参数开启时会运行 Java 8 测试, 尽管 `-DskipTests` 配置项开启时也会运行。要运行这些测试系统必须已安装 Java 8。如果已安装 JDK 8 但又不是系统默认 JDK, 可以运行测试前将 `JAVA_HOME` 设置成 JDK 8。

10. 基于 Yarn 编译 PySpark (Building for PySpark on YARN)

如果使用 Maven 编译 jar, 则只支持 PySpark on YARN。另外, 基于 Red Hat 内核的系统中, 使用这个集成包编译会有一个问题, (详见 [SPARK-1753](#))。如果你想在装有 Red Hat 操作系统的 Yarn 集群上运行 PySpark, 我们建议在其他地方编译 Jar 包, 然后封装到集群中。目前我们正在定位这个问题的确切原因。

11. 不使用 Yarn 的 Hadoop 依赖库打包

通过 `mvn package` 命令生成的 jar 包默认将包括所有的 Spark 依赖库, 包括 Hadoop 以及某些 Hadoop 体系 (Hadoop-ecosystem) 的工程。采用 Yarn 方式部署会导致在 `executor classpath` 出现多个不同版本的 Jar 包: 即每个节点包括 `yarn.application.classpath` 参数。使用 `hadoop-provided` 配置参数编译可以不集成 Hadoop 体系的工程, 例如 ZooKeeper 和 Hadoop 本身。

■ 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、商业案例探讨！

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