亚太研究院

Spark 官方文档翻译

文档首页（V1.1.0）

翻译者　颜军

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

文档首页（V1.1.0）

（翻译者：颜军）

Apache Spark 原文档链接：
http://spark.apache.org/
http://spark.apache.org/docs/latest/
http://spark.apache.org/downloads.html

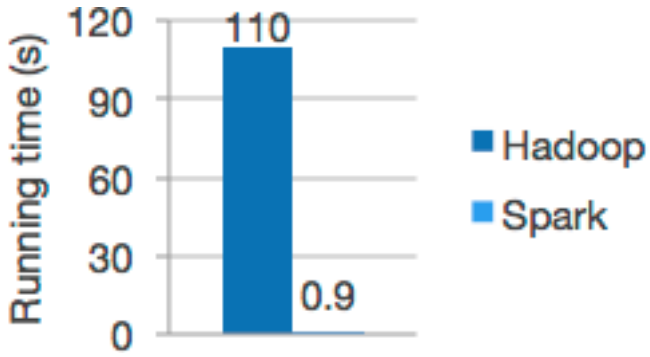
目录

Apache Spark™ 是一个高效的通用的大规模数据处理引擎。

速度

比内存中的 Hadoop MapReduce 快 10 倍 比硬盘上的 Hadoop MapReduce 快 100 倍

Spark 有一个高级的 DAG 执行引擎，支持循环迭代的数据流和内存中的计算。



Hadoop 和 Spark 中的逻辑回归（Logistic regression）

易用

用 Java、Scala 或者 Python 迅速开发应用。

Spark 提供超过 80 种高层级操作，使得容易构建并行应用。 你也可以通过 Scala 和 Python 命令行交互地使用它。

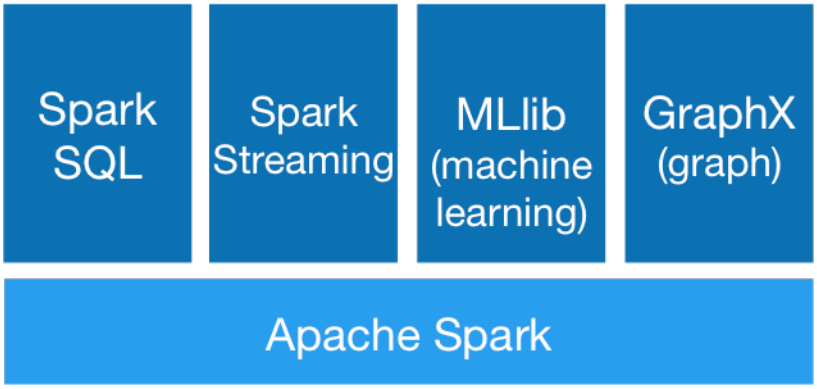
```
file = spark.textFile("hdfs://...")

file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

通过 Python API 的 Spark 单词计数

通用

结合 SQL、流（streaming）和复杂性分析（complex analytics)。

Spark提供一套高层工具栈 ,包括 Spark SQL ,机器学习的MLlib， GraphX ,和 Spark Streaming。 你可以在一个应用中无缝地结合这些框架。



集成 Hadoop

Spark 可以运行在 Hadoop 2 的 YARN 集群之上，可以读取任何现存 Hadoop 数据。

如果你已有一个Hadoop 2 集群，无需任何安装即可运行Spark。 此外，Spark也很容易独立运行， 或者在EC2或Mesos之上。 它可以读取HDFS, HBase, Cassandra, 以及任何Hadoop数据源。

社区

Spark被众多组织采用处理大数据集。 你可以在 Spark Summit 峰会或者 Powered By 网页找到案例。

有很多途径接触社区：

- 使用邮件列表 提问
- 参加Bay Area Spark meetup 和 Spark Summit.
- 我们使用JIRA 进行开发管理

贡献者

Apache Spark 由超过 50 位的众多开发者构建。 自从 2009 年项目启动，超过 250 为开发者为 Spark 做出贡献！

项目的 提交者（commiters） 来自 12 个组织。

如果你愿意参与Spark或者其上的库，学习如何贡献.

开始

无论你有 Java 或 Python 背景，学习 Spark 都很容易：

- 下载最新版本，可以在本地笔记本上运行。

- 阅读快速开始指南.
- Spark Summit 2013 包含免费的 视频 和 习题，你可以在Amazon EC2 上运行。
- 学习如何布署 Spark集群。

下载 Spark

Spark最新版本是Spark 1.1.0, 发行于 2014 年 9 月 11 日 (release notes) (git tag)

1. 选择 Spark 版本：
2. 选择包类型：
3. 选择下载类型：
4. 下载 Spark: spark-1.1.0.tgz
5. 使用1.1.0 signatures and checksums校验下载版本。

连接到 Spark

Spark artifacts托管在Maven Central。你可以用以下参数添加Maven依赖:
groupId: org.apache.spark
artifactId: spark-core_2.10
version: 1.1.0

开发和维护分支

如果你对正在开发的代码有兴趣或者想做出贡献，可以从 Git 获取主分支：

```
# Master development branch
git clone git://github.com/apache/spark.git

# 1.1 maintenance branch with stability fixes on top of Spark 1.1.0
git clone git://github.com/apache/spark.git -b branch-1.1
```

下载Spark之后，你可以从文档页找到安装和编译说明。

全部版本

- Spark 1.1.0 (Sep 11 2014)
- Spark 1.0.2 (Aug 05 2014)
- Spark 1.0.1 (Jul 11 2014)
- Spark 1.0.0 (May 30 2014)
- Spark 0.9.2 (Jul 23 2014)
- Spark 0.9.1 (Apr 09 2014)
- Spark 0.9.0 (Feb 02 2014)

- Spark 0.8.1 (Dec 19 2013)
- Spark 0.8.0 (Sep 25 2013)
- Spark 0.7.3 (Jul 16 2013)
- Spark 0.7.2 (Feb 06 2013)
- Spark 0.7.0 (Feb 27 2013)
- Spark 1.1.0 (Sep 11 2014)
- Spark 1.0.2 (Aug 05 2014)
- Spark 1.0.1 (Jul 11 2014)
- Spark 1.0.0 (May 30 2014)
- Spark 0.9.2 (Jul 23 2014)
- Spark 0.9.1 (Apr 09 2014)
- Spark 0.9.0 (Feb 02 2014)
- Spark 0.8.1 (Dec 19 2013)
- Spark 0.8.0 (Sep 25 2013)
- Spark 0.7.3 (Jul 16 2013)
- Spark 0.7.2 (Feb 06 2013)
- Spark 0.7.0 (Feb 27 2013)

Spark 概览

Apache Spark是一个高效的通用的集群计算系统。 它提供高层级的Java, Scala 和 Python 接口，和优化的通用图计算引擎。 同时支持丰富的高级工具集，如处理SQL和结构化数据的 Spark SQL ，机器学习的 MLlib，图处理的 GraphX ，和 Spark Streaming。

下载

从项目网站的 下载页获取Spark。 这份文档是关于Spark 1.1.0 版本的。下载页包括适配许多常见 HDFS 版本的Spark。如果你愿意从头编译Spark，访问 building Spark with Maven。

Spark同时运行在Windows 和 类UNIX 系统 (如 Linux, Mac OS)上。很容易单机运行——你只需要安装 java并配置 PATH或 JAVA_HOME环境变量指向安装目录。

Spark运行在 Java 6+ 和 Python 2.6+上。对于 Scala ,Spark 1.1.0使用 Scala 2.10。你需要使用兼容版本的 Scala (2.10.x)。

运行示例和命令行

Spark带有一些示例程序。Scala, Java and Python 例子在 examples/src/main 目录下。要运行 Java 或 Scala 示例程序，在Spark顶级目录使用 bin/run-example <class> [params] (实际上，它调用更加通用的 spark-submit 脚本来启动程序)。例如

./bin/run-example SparkPi 10

你也可以通过一个修改过的 Scala shell 来交互地运行 Spark，这是一个学习框架的好方法。

./bin/spark-shell --master local[2]

使用 --master参数指定 一个分部式系统的主URL，或者 local指定本地单线程运行，或 local[N]指定本地N线程。开始你应该使用 local来试用。查看完整参数表，使用 --help 参数。

Spark也提供Python API。若要通过Python解释器交互运行Spark，使用 bin/pyspark：

./bin/pyspark --master local[2]

也有 Python 下的示例程序，例如

./bin/spark-submit examples/src/main/python/pi.py 10

运行一个集群

Spark 集群模式概览介绍了运行集群的一些重要概念。 Spark既可以独立运行，也可以运行在既存的集群之上。它现在提供了几种布署选项：

- Amazon EC2：我们的EC2 脚本让你 5 分钟启动一个集群
- 独立布署模式：在私有群上布署Spark的最简单途径
- Apache Mesos
- Hadoop YARN

接下来

编程指南：

- 快速开始：Spark API简要介绍；开始！
- Spark 编程指南： 详细浏览Spark所有支持语言(Scala, Java, Python)
- Spark 上的模块：
  - Spark Streaming：处理实时数据流
  - Spark SQL：支持结构化数据和关系型查询
  - MLlib：内置机器学习库
  - GraphX：Spark的新图处理API
  - Bagel (Pregel on Spark)：更早的简单的图处理模型

API文档：

- Spark Scala API (Scaladoc)
- Spark Java API (Javadoc)
- Spark Python API (Epydoc)

布署指南：

- 集群概览：集群上的概念和组件简介
- 提交程序：程序的打包和布署
- 布署模式：
  - Amazon EC2：让你在 5 分钟在EC2 启动一个集群的脚本
  - 独立布署模式：无需第三方集群管理工具快速运行一个独立集群
  - Mesos：使用 Apache Mesos布署一个私有集群
  - YARN：在Hadoop NextGen (YARN)上布署Spark

其他文档：

- 配置：通过配置系统定制化Spark
- 监控：追踪你程序的行为
- 调优指南：优化性能和内存使用的最佳实践
- 排程：在程序内和程序间配置资源
- 安全： Spark安全支持
- 硬件要求：集群硬件推荐配置
- 第三方Hadoop发行版：使用常见Hadoop发行版
- 与其他存储系统集成：
  - OpenStack Swift
- 用Maven编译Spark：用Maven系统编译Spark
- 为Spark做贡献

外部资源：

- Spark主页
- 邮件列表：在这提关于Spark问题
- AMP Camps：在UC Berkeley的一系列训练营，关于Spark、Spark Streaming、Mesos以及其他方面的交流和练习。有许多在线的免费 视频, 讲稿和 练习。
- 代码示例：Spark的 examples子目录下有更多(Scala, Java, Python)

社区

注册 用户邮件列表来获取关于使用Spark的帮助及跟进Spark的开发。

如果你在San Francisco Bay Area，每隔几周有一个定期的 Spark meetup，过来可以见到开发者和其他用户。

最后，如果你愿意为Spark贡献代码，阅读 how to contribute。

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