亚太研究院

Spark 官方文档翻译

EC2 (V1.1.0)

翻译者　樊登贵

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

EC21.1.0

（翻译者：樊登贵）

Running Spark on EC2，原文档链接：

http://spark.apache.org/docs/latest/ec2-scripts.html

目录

第一章 Spark on EC2 运行模式

http://spark.apache.org/docs/latest/ec2-scripts.html

Spark的 ec2 目录下的 spark-ec2 脚本 ， 可以用来启动、管理和关闭Amazon EC2（亚马逊弹性云计算）中的Spark集群， 并且会为你在集群上自动装配Spark、Shark和HDFS等计算框架。 本指南介绍了如何使用 spark-ec2 启动集群，如何在集群上运行任务（jobs）， 以及如何将集群关闭等。 在使用Amazon EC2 服务前，需要在 亚马逊网络服务网站（AWSs）注册一个EC2 帐号。

脚本 spark-ec2 是用来管理多个名称集群的， 你可以启动一个新的集群（告诉脚本集群规模并为集群取名）， 关闭现有的群集， 或者登录到一个特定的集群。 每个集群通过将其所有节点放入 EC2 安全组的方式来进行识别， 安全组的名称是从集群的名称派生来的。例如，一个名为 test 的集群将包含一个位于 test-master 安全组的 master 节点和多个位于 test-slaves 安全组的 slave 节点。 spark-ec2 将根据你所请求的群集的名称来创建这些安全组， 反之，你也可以使用这些安全组来识别位于 Amazon EC2 控制台中每个集群的所有节点。

1.1 准备工作

- 创建Amazon EC2 密钥对：通过Amazon Web Services (AWS) 控制台登录到你的AWS帐号 ,点击左侧栏的密钥对 ,创建和下载一个密钥；确保私钥文件的设置权限为 600（即仅本人可读写）， 这样 ssh 才能 起作用。
- 设置Amazon EC2 的环境变量： 使用 spark-ec2前， 需要从 AWS页面， 通过 “账号>安全凭证>访问凭据”的方式， 来设置 Amazon EC2 访问密钥 ID 的环境变量 AWS_ACCESS_KEY_ID， 及其相关密钥的环境变量 AWS_SECRET_ACCESS_KEY

1.2. 启动集群

- 进入Spark目录 ec2
- 运行 ./spark-ec2 -k <keypair> -i <key-file> -s <num-slaves> launch <cluster-name> ，其中 <keypair> 是EC2 密钥对的名称（创建时已给定）， <key-file> 是密钥对的私钥文件， <num-slaves> 是要启动的slave 节点数，（首次启动可取为 1），以及 <cluster-name> 为集群的名称。

- 全部启动之后，检查集群的调度器是否已经启动，并从 web UI 查看所有的 slave 节点，它们将被打印在脚本（通常为 `http://<master-hostname>:8080` ）的末尾处

你还可以运行 `./spark-ec2 --help` 查看到更多有关使用选项的帮助信息，其中下列选项中需要注意：

- `--instance-type=<INSTANCE_TYPE>` 指定一个EC2 实例类型。目前，该脚本只支持 64 位的实例类型，默认类型是`m1.large`（其中包含 2 个内核和 7.5 GB RAM）。有关更多实例类型的信息，请参阅亚马逊网页的 EC2 实例类型和 EC2 定价。
- `--region=<EC2_REGION>` 指定一个可配置EC2 实例的地区， 默认地区是美国东部地区，即 `us-east-1` .
- `--zone=<EC2_ZONE>` 指定 EC2 实例的一个可用性区域（availability zone）； 如果指定的这个可用性区域的 EBS（Elastic Block Storage，弹性块存储）容量不足，系统就会返回一个错误信息，这时你应该尝试另一个可用性区域来重建 EC2 实例。
- `--ebs-vol-size=GB` 附加一个指定大小的EBS卷到每个节点上， 以便在集群重启时，所有节点都拥有一个持久的HDFS集群（见下文）。
- `--spot-price=PRICE` 启动worker节点作为 现货实例，（Spot Instances, 亚马逊的现买现卖业务），竞标最高价格（以美元计）。
- `--spark-version=VERSION` 预加载指定版本的Spark集群， 其中 VERSION 可以是一个版本号（例如 “0.7.3”）或特定的git哈希值（默认为当前版本号）。
- 如果某次启动失败，例如缺少私钥文件的访问权限，可以运行`launch --resume` 在当前的集群上重新启动设置过程。

1.3. 运行 Applications

- 进入Spark目录 ec2
- 运行 ./spark-ec2 -k <keypair> -i <key-file> login <cluster-name> SSH 免密码登陆集群，其中 <keypair> 和 <key-file> 同上。(方便起见，你还可以使用 EC2 控制台)
- 为了在集群上部署代码或数据，你可以登录并使用提供的脚本 ~/spark-ec2/copy-dir，如果给定了目标路径，它将会通过 RSYNC 同步备份到所有 slave 上的同一位置。
- 如果你的 application 需要访问大型数据库， 最快的方法是在 Amazon S3（亚马逊简易存储服务）或亚马逊的 EBS 设备上加载这些数据库到节点的 HDFS 实例中。spark-ec2 已经创建了一个 HDFS 实例， 安装在目录/root/ephemeral-hdfs 下， 并且可

以使用该目录中的脚本 bin/hadoop 来访问。 注意，当你停止和重启机器时， 位于 HDFS 中的数据将会消失。

- 在目录 /root/persistent-hdfs 下还有一个*持久的 HDFS* 实例， 集群重启时它将会保存数据以防丢失。 通常，每个节点都有相对较小（约 3 GB）的持久性数据存储空间，但可以使用 spark-ec2 中的 --ebs-vol-size 选项，为每个节点附加一个持久的 EBS 卷，用来存储持久的 HDFS 数据。
- 最后， 如果运行 application 时出现错误， 可以从调度器的工作目录 (/root/spark/work) 下找到你的 application，并查看它的 slave 日志; 你还可以通过 web UI (http://<master-hostname>:8080) 查看集群的状态 。

1.4. 配置

你可以编辑 /root/spark/conf/spark-env.sh 来设置Spark的配置选项，如JVM选项，这个文件需要复制到每一台机器上来显示更改信息。 这样做的最简单方法就是使用我们提供的称为copy-dir的脚本， 首先在master上编辑文件spark-env.sh， 然后运行 ~/spark-ec2/copy-dir /root/spark/conf 将其同步备份到所有的worker上。

配置指南中给出了可用的配置选项。

1.5. 终止集群

注意， 如果关闭EC2 上的节点后没有办法恢复它的数据， 请确保在停止之前拷贝所有重要的信息。

- 进入Spark目录 ec2
- 运行 ./spark-ec2 destroy <cluster-name>

1.6. 暂停和重启集群

spark-ec2 还支持用户暂停一个集群。 在这种情况下，虚拟机停止运行但并未终止运行，所以尽管它们 ***丧失了暂存磁盘上的所有数据*** ，但却把数据保存在它们的根分区和 persistent-hdfs中 。 停止的机器不需要EC2 的任何运行成本， 但 ***仍需*** EBS的存储成本。

- 中止集群： 进入 ec2 目录，然后运行 ./spark-ec2 stop <cluster-name>
- 重启集群： 运行 ./spark-ec2 -i <key-file> start <cluster-name>

- 最终摧毁集群，中止消耗 EBS 空间：运行 ./spark-ec2 destroy <cluster-name>，如上一节所述。

1.7. 局限性

- 支持 “集群计算” 的节点是有限的——虽然没有办法指定一个安全组，但是可以在 <clusterName>-slaves 组中手动启动 slave 节点，然后使用 spark-ec2 launch --resume 启动群集。

如果有任何关于局限性的补充和建议，随时欢迎 提出宝贵意见！

1.8. 访问 S3 中数据

Spark 的文件接口层使其能够使用和Hadoop同样的URI格式处理Amazon S3 中的数据。首先需要在S3 中指定一个URI格式的路径： s3n://<bucket>/path 作为输入 ， 还需要在程序之前设置环境变量 AWS_ACCESS_KEY_ID 和 AWS_SECRET_ACCESS_KEY， 或通过 SparkContext.hadoopConfiguration 来设置你的Amazon安全凭证。 使用Hadoop输入库来访问S3 的详细说明可参阅 Hadoop S3 页面。

除了使用单输入文件外， 你还可以通过简单地给目录指定路径来使用目录下的所有文件作为输入。

Spark 亚太研究院

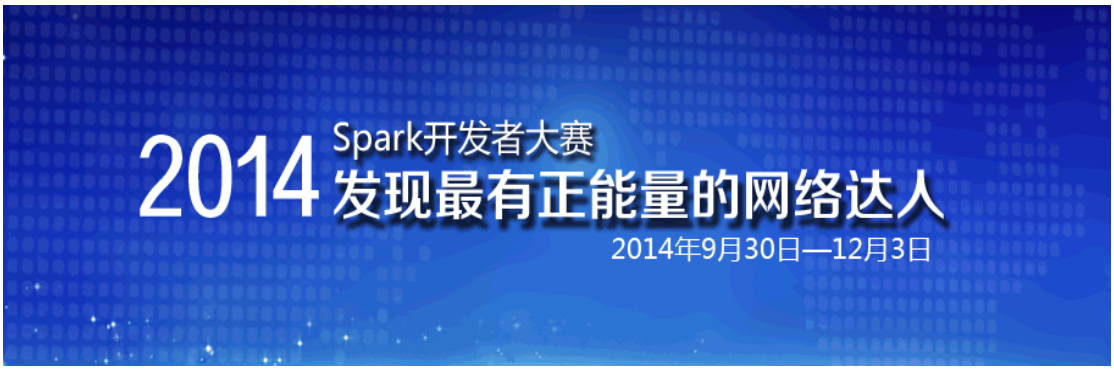
Spark 亚太研究院是中国最专业的一站式大数据 Spark 解决方案供应商和高品质大数据企业级完整培训与服务供应商，以帮助企业规划、架构、部署、开发、培训和使用 Spark 为核心，同时提供 Spark 源码研究和应用技术训练。针对具体 Spark 项目，提供完整而彻底的解决方案。包括 Spark 一站式项目解决方案、Spark 一站式项目实施方案及 Spark 一体化顾问服务。

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