亚太研究院

Spark 官方文档翻译

Spark 作业调度

(v1.1.0)

翻译者　王庆刚

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

Spark 作业调度

(v1.1.0)

（翻译者：王庆刚）

Job Scheduling，原文档链接：

http://spark.apache.org/docs/latest/job-scheduling.html

目录

1. 概述.....6

2. 应用间的调度.....6

3. 应用内的调度.....7

4. 公平调度池.....7

4.1 池的默认行为.....8

4.2 配置池的属性.....8

1. 概述

Spark 有很多措施能够使得它在计算过程中进行资源调度。首先，就像集群模式概述中讲的那样，每一个 spark 应用(即 sparkContext 的实例)运行一个独立的执行进程。Spark 的集群管理器为跨应用的调度提供处理措施。其次，在每一个 spark 的应用中，不同的线程可能会导致多个作业（spark actions）的同时 运行。这一点非常常见如果你的应用跨网络提交请求；比如，shark 的服务端就是这样工作。Spark 在每个 sparkContext 包含了一个公正的调度策略来进行资源调度。

2. 应用间的调度

当 Spark 以集群模式运行的时候，每一个 spark 的应用都会获得一个独立的 jvm 使其可以执行作业跟保存相关数据。如果有多个用户需要共同使用你的集群，Spark 会根据集群的管理器具有多种策略来进行资源分配。

对所有的集群管理器来说最简单的策略就是静态的资源分配。在这种方法中，每一个应用将会被给予系统能够分配的最大的资源并且在整个运行周期内占有这些资源。一下将会介绍 spark 中的 standalone 模式，yarn 模式以及 coarse-grained memos 模式。

Standalone 模式：默认的，当应用被提交到 standalone 模式下，集群会进入 FIFO 排序模式（先进先出），并且每一个应用都会尝试去使用所有可用的节点。你可以通过使用 spark.cores.max 或者修改默认的参数让应用不再使用 spark.deploy.defaultCores 这个参数来设置从而来限制节点的数量。 Spark 除了能够设置 cpu 的核数之外，还可以通过 spark.executor.memory 这个参数来设置应用所使用能够使用的内存

Mesos 模式：为了使用 Mesos 的静态分区，可以通过设置 spark.mesos.coarse 这个属性为 true，并且有选择性的设置 spark.cores.max 来限制每个应用在所使用的资源 ,就像 standalone 模式一样。另外还可以通过设置 spark.executor.memory 这个属性来控制执行器的内存使用

YARN 模式：可以通过配置 --num-executors 这个属性使 spark 的 YARN 客户端控制集群下执行器的数量，可以使用-executor-memory 与 --executor-cores 这两个属性来限制每个执行器使用的资源。

Mesos 模式下还存在一个额外的可以动态分配 CPU 核数选项。在这种模式下，每一个应用仍然会有一个固定并且独立的内存（通过 spark.executor.memory 设置），但是当这个应用不在运行作业的实时，别的应用可以使用这些 cpu。这种模式尤为有效当有大量非活跃的应用的时候，比如单独用户的 shell 会话。然而，这种模式带有不可预测的延时风险，因为当一个应用又开始运行的时候 ,前面的设置会导致这个应用需要一段时间才能够重新获

得它原有的资源。想用这个模式的话，直接用 mesos:// URL 并不用设置 spark.mesos.coarse 为 true。

需要指出的是上面的模式中没有任何一种可以在众多应用中共享内存。如果你想共享内存，我们建议运行一个单独的服务器的应用使其可以通过查询相同的 RDDs 能够提供多个请求。像 Shark 的 JDBC 服务端就是通过这种方式来提供 SQL 的查询。在将来的版本中，基于内存的存储系统比如 Tachyon 将会提供其他的方式来共享 RDDs

3. 应用内的调度

在一个给定的 spark 应用（即 SparkContext 的实例）中，多个平行的作业可以同时运行如果这些作业由不同的线程提交的话。在这一节中的“作业”，我们称之为一个 spark 的 action（比如 save 或者 collect 操作）与任何需要运行来评估 action 的任务。Spark 的调度器是完全线程完全的并且可以支持应用为多个请求提供服务。

默认的，Spark 的调度器使用 FIFO 模式运行作业。每一个作业都会被分割成多个“阶段”（例如 map， reduce），并且第一个的作业在其阶段中存在任务运行的期间会优先获取所有可以获得的资源，然后是第二个作业获得优先权，以此类推。如果排在作业调度队列最前面的作业不需要使用整个集群，之后的作业可以立即启动，但是如果最前面的作业需要占用绝大多数资源，那么接下来的作业将会明显的被耽搁。

在 Spark 的 0.8 版本之后，可以通过配置来使不同的作业之间能够公平共享资源，在这种模式下，Spark 将不同的作业的任务放置到一个循环模式下，从而使得所有的作用能够获得相对平均的集群资源。这就意味着短的作业可以在长任务运行的时候立马启动并能够获得一个非常好响应时间而不用去等待长任务去完成。这个模式对于多个用户来说是最好的设置方案。

为了能够使用公平调度器，在配置 SparkContext 的时候可以通过设置 spark.scheduler.mode 为 fair

```
val conf = new SparkConf().setMaster(...).setAppName(...)conf.set("spark.scheduler.mode", "FAIR")val sc = new SparkContext(conf)
```

4. 公平调度池

公平调度也能够支持分组的任务到池中，并且为每一个池分配不同的调度选项(例如 权重)。这在为一些非常重要的作业生成一个优先级高的池的时候非常有用，或者是把一个用户的作业集中起来并分组通过用户来分配资源而不管这个用户具有多少正在运行的作业。这种方式是 Hadoop Fair scheduler 的仿照。

在没有任何干预的情况下，新提交的作业将会进入到默认的池中，但是作业的池可以通过在 Spark Context 的线程中那个把 spark.scheduler.pool 添加到“local property”中并提交。可以按如下的方式做

```
// Assuming sc is your SparkContext variablesc.setLocalProperty("spark.scheduler.pool", "pool1")
```

在设置了这个属性之后，所有通过这个线程（线程中的 RDD.save count collect 等等操作）提交的作业将会使用这个池的名字。这种设置是基于线程的这使得一个线程能够代表一个用户运行很多的作业变得非常方便。如果你想清除某个线程所关联的池，按照如下调用就可以

```
sc.setLocalProperty("spark.scheduler.pool", null)
```

4.1 池的默认行为

默认的，每一个池对于集群来说都是公平的（跟默认池中每一个作业对应集群是公平的一样），但是在每一个池中，作业是按照 FIFO 排序运行。比如，如果你想为每一个用户建立一个池，这意味着每一个用户对应集群来说将会等同对待，每一个用户的查询都会按照顺序执行而不是之后的查询获取这个用户以前查询的资源

4.2 配置池的属性

池的某些特定的属性可以通过配置文件来进行修改，每一个池支持一下三种属性：

- schedulingMode: 这个属性可以为FIFO或者是FAIR，来控制池中队列里面的作业是先出的优先获取资源还是公平的共享资源
- weight:这个可以设置集群中不同池之间的权重，默认的，所有的池权重都为 1，如果你把某个特定的池的权重设置为 2 ,他相对与其他的池来说将会获得 2 倍的资源。如果设置一个很高的值例如 1000 ,这使得这个池能够相对应其他的池总是能够优先获取资源，这个权重为 1000 的池总是能够在它的作业活跃的时候执行任务
- minShare: 除了大体上的设置，每一个池可以给予一个最小的分配（例如cpu的核数）按照管理员的意愿。公平的调度器总是会在先满足所有活跃的池的最小的分配值然后重新通过权重重新分配剩余的资源。这个minShare属性是总是能够获取一部分资源的另外一种方式而不用通过设置一个相对于其他作业非常他的权重。默认的每一个池的minShare这个属性为 0

池的属性可以通过生成一个类似于 conf/fairscheduler.xml.template 的 xml，或者是在 SparkConf 中设置 spark.scheduler.allocation.file

```
conf.set("spark.scheduler.allocation.file", "/path/to/file")
```

这个 XML 的格式为每一个池都会对应一个 <pool> 的元素 ,在里面有不同的元素对应于不同的设置，例如

```
<?xml version="1.0"?><allocations>  <pool name="production">
<schedulingMode>FAIR</schedulingMode>  <weight>1</weight>
<minShare>2</minShare>  </pool>  <pool name="test">
<schedulingMode>FIFO</schedulingMode>  <weight>2</weight>
<minShare>3</minShare>  </pool></allocations>
```

在 conf/fairscheduler.xml.template 中有一个完全的例子。需要指出的是不在 xml 文件中配置的任何池的多有设置都会为默认属性（scheduling mode FIFO, weight 1, and minShare 0）

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