


vincent_hv

Talk is cheap, show the code!

博客园 闪存 首页 新随笔 联系 管理 订阅 

随笔- 86 文章- 0 评论- 3

昵称：vincent_hv

园龄：10个月

粉丝：7

关注：1

[+加关注](#)

【转】Spark 体系结构

原文地址：<http://jerryshao.me/architecture/2013/03/29/spark-overview/>

援引@JerryLead的系统架构图作为Spark整体结构的一个 birdview：

<	2013年10月					
日	一	二	三	四	五	六
29	30	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

搜索

<input type="text"/>	<input type="button" value="找找看"/>
<input type="text"/>	<input type="button" value="谷歌搜索"/>

常用链接

[我的随笔](#)

[我的评论](#)

[我的参与](#)

[最新评论](#)

[我的标签](#)

[更多链接](#)

最新随笔

1. [linux解压zip乱码解决方案](#)
2. [全能系统监控工具dstat](#)
3. [【转】linux sar命令详解](#)
4. [【原】gnome3增加自定义程序快捷方式](#)
5. [【原】Ubuntu13.04安装、卸载Gnome3.8](#)
6. [【原】安装、卸载、查看软件时常用命令](#)
7. [【原】中文Ubuntu主目录下的文档文件夹改回英文](#)
8. [【原】Ubuntu ATI/Intel双显卡 驱动装](#)
9. [【原】Ubuntu 12.04 ATI显卡设置双屏显示](#)
10. [【转】Hadoop vs Spark性能对比](#)

随笔分类

Android(8)
Hadoop(2)
Java(20)
JVM(3)
Linux(23)
others(1)
Scala(5)
Spark(20)
数据结构与算法(2)

积分与排名

积分 - 5935
排名 - 17402

最新评论

1. Re: 全能系统监控工具dstat
感觉好高级的样子，我也下载来玩玩
--花灞奶
2. Re: 【原】Ubuntu13.04安装、卸载Gnome3.8
马上应该有13.10了。
--杨
3. Re: scala实现kmeans算法
在oschina上一位大牛给我的指点，原文上，供跟多的孩纸学习：oldpig 发表于 2013-09-03 10:45 1. Source.getLines() 返回的Iterator已经够用了，不需要toArray 2. 随机初始化k个质心，可以考虑使用ray.fill 3. 如果你要测算法的计算时间，l将两条println语句放到startTime之前 4 计算movement可以考虑使用...
--vincent_hv

阅读排行榜

1. Ubuntu 13.04 完全配置(3095)
2. Android控件TextView的实现原理分析(213)
3. 【转】JVM (Java虚拟机) 优化大全案例实战(175)
4. 【转】Spark：一个高效的分布式计算系统(139)
5. 修改Ubuntu12.04 开机启动菜单，包括系统启动等待时间，系统启动顺序(13)

评论排行榜

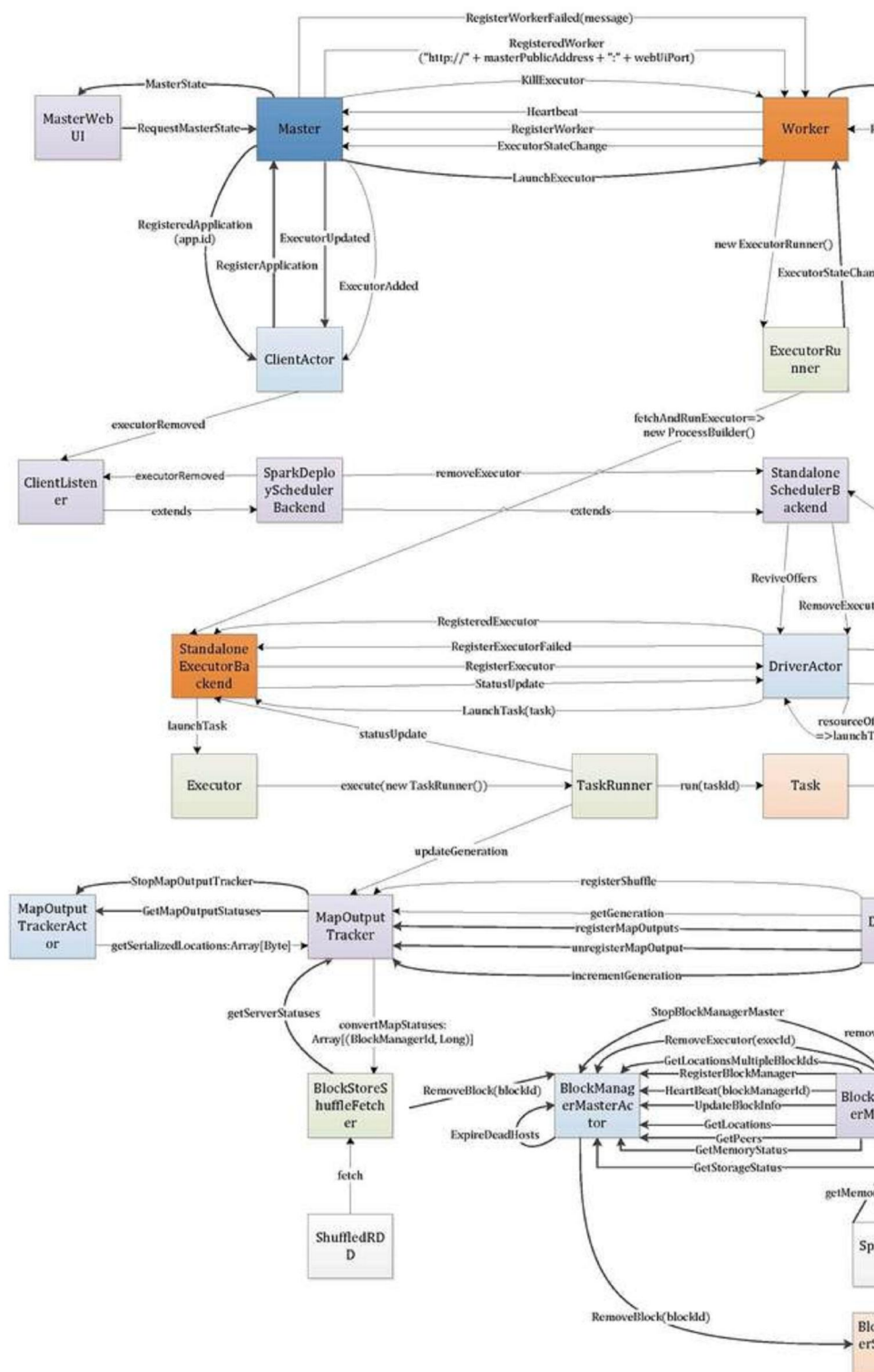
1. 【原】Ubuntu13.04安装、卸载Gnome3.8(1)
2. scala实现kmeans算法(1)
3. 全能系统监控工具dstat(1)
4. 【转】linux 杀命令详解(0)
5. 【原】gnome3增加自定义程序快捷方式(0)

推荐排行榜

1. 【转】Spark源码分析之-Storage模块(2)
2. 【转】弹性分布式数据集：一种基于内存的集群计算的容错性抽象方法(1)
3. 【转】Spark：一个高效的分布式计算系统(1)
4. linux解压zip乱码解决方案(1)
5. 全能系统监控工具dstat(1)

Architecture of Spark-0

Drawn by JerryLead
csxuljie@gmail.com



整体上Spark分为以下几个主要的子模块：

- **deploy**: deploy模块包括Master, Work和Client，参见architecture图的最上部分。deploy主要负责启动和调度用户实现的Spark application并且分配资源给用户 application，类似于Hadoop YARN框架。
- **scheduler**: scheduler主要负责调度用户application内的tasks，根据部署方式的不同Spark实现了多种不同的scheduler，包括LocalScheduler，ClusterScheduler等。

- rdd: rdd类似于一个分布式的数据集，用户可以根据rdd所提供的api进行数据集的 操作，rdd模块是用户交互的主要模块。
- storage: storage模块主要负责数据集，也就是rdd的存取。根据设定的不同，数 据可以保存在内存、磁盘或是两者。Spark与Hadoop MapReduce最大的不同在于MapReduce 将数据保存在HDFS上，而Spark则由自己的存储系统。

当然还有一些其他的子模块，可以参考上图。

Spark采用了Actor的设计方式，整体架构，包括各子模块的设计上都是采用master-slave模 式，master和slave之间通信的主要协议可以参见上图。

分类: [Spark](#)

绿色通道： [好文要顶](#) [关注我](#) [收藏该文](#) [与我联系](#) 



[vincent_hv](#)

[关注 - 1](#)

[粉丝 - 7](#)

[+加关注](#)

0

0

(请您对文章做出评价)

« 上一篇: [【转】Spark快速入门指南](#)

» 下一篇: [【译】Spark官方文档——Spark Configuration \(Spark配置 \)](#)

posted @ 2013-09-10 17:19 [vincent_hv](#) 阅读(69) 评论(0) [编辑](#) [收藏](#)

[刷新评论](#) [刷新页面](#) [返回顶部](#)

注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

[博客园首页](#) [博问](#) [新闻](#) [闪存](#) [程序员招聘](#) [知识库](#)

最新IT新闻:

- [Google更新reCAPTCHA验证码，降低对人类的难度](#)
 - [互联网档案馆默认启用HTTPS](#)
 - [转基因鲑鱼有望在美上市](#)
 - [惠普起诉东芝三星等操纵光驱价格 要求三倍赔偿](#)
 - [传易信接洽联通移动 或打通三网流量费用全免](#)
- » [更多新闻...](#)

最新知识库文章:

- [软件开发启示录——迟到的领悟](#)
 - [《黑客帝国》里的锡安是不是虚拟世界](#)
 - [深入理解Linux中内存管理](#)
 - [工程师文化引出的组织行为话题](#)
 - [如何用美剧真正提升你的英语水平](#)
- » [更多知识库文章...](#)

Copyright ©2013 [vincent_hv](#)