

vincent_hv

Talk is cheap, show the code!

博客园 闪存 首页 新随笔 联系 管理 订阅 XML

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

【原】Spark 编程指南

尊重原创，注重版权，转贴请注明原文地址：<http://www.cnblogs.com/vincent-hv/p/3322966.html>

1、配置程序使用资源：

```
System.setProperty("spark.executor.memory", "512m")
```

2、创建自己的SparkContext对象：

```
val sc = new SparkContext("master", "Job name", "$SPARK_HOME", "Job JARs")
```

3、创建RDDs

```
sc.parallelize(List(1, 2, 3))    // 将scala原生的集合转换成RDD
sc.textFile("directory/*.txt")    // 将本地磁盘上的文本转换成RDD
sc.textFile("hdfs://namenode:port/path/file")    // 将分布式文件系统上的文件转换成RDD
```

4、基本的RDD Transformations操作：

```
val nums = sc.parallelize(List(1, 2, 3))
val squares = nums.map(x => x * x)    // {1, 4, 9}
val even = squares.filter(_ % 2 == 0)    // {4}
nums.flatMap(x => 1 to x)    // {1, 1, 2, 1, 2, 3}
```

5、基本的Actions操作：

```
val nums = sc.parallelize(List(1, 2, 3))
nums.collect()    // Array(1, 2, 3)
nums.take(2)    // Array(1, 2)
nums.count()    // 3
nums.reduce(_ + _)    // 6
nums.saveAsTextFile("hdfs://file.txt")
```

6、针对Key-Value对的作业：

```
val pair = (a, b)
pair._1    // a
pair._2    // b
```

7、一些Key-Value对的操作：

```
val pets = sc.parallelize(List(("cat", 1), ("dog", 1), ("cat", 2)))
pets.reduceByKey(_ + _)    // {(cat, 3), (dog, 1)}
pets.groupByKey()    // {(cat, Seq(1, 2)), (dog, 1)}
pets.sortByKey()    // {(cat, 1), (cat, 2), (dog, 1)}
```

8、其他Key-Value的操作：



```
val visits = sc.parallelize(List(("index.html", "1.2.3.4"), ("about.html", "3.4.5.6"), ("index.html", 1.3.3.1)))
val pageNames = sc.parallelize(List(("index.html", "Home"), ("about.html", "About")))
visits.join(pageNames)    // ("index.html", ("1.2.3.4", "Hmoe"))
```

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

搜索

<input type="text"/>	找找看
<input type="text"/>	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、查看软件时常用的命令
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 驱动安装
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)

```
// ("index.html", ("1.3.3.1", "Home"))
// ("about.html", ("3.4.5.6", "About"))
visits.cogroup(pageNames) // ("index.html", (Seq("1.2.3.4", 1.3.3.1), Seq("Home")))
                           // ("about.html", (Seq("3.4.5.6"), Seq("About")))
```



9、控制Reduce Tasks的数量：

所有的RDD组操作都可以选择设置第二个参数来控制tasks的数量

```
words.reduceByKey(_ + _, 5)
words.groupByKey(5)
visits.join(pageViews, 5)
```

当然，也可以通过设置spark.default.parallelism属性值来控制

10、使用本地变量：

在闭包中你使用任何外部变量都将自动的传递到集群：

```
val query = Console.readLine()
pages.filter(_ contains(query)).count()
```

一些注意事项：

- Each task gets a new copy(updates aren't sent back)
- 变量必须序列化
- 不要使用一个对象的外部域

11、集群中有较大危害的示例：



12、其他RDD操作：

- sample()：确定样本子集
- union()：合并两个RDD
- cartesian()：交叉乘积
- pipe()：通过外部程序

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性能测试报告](#)

» 下一篇：[【转】Spark：一个高效的分布式计算系统](#)

posted @ 2013-09-15 19:17 [vincent_hv](#) 阅读(109) 评论(0) [编辑](#) [收藏](#)

[刷新评论](#) [刷新页面](#) [返回顶部](#)

积分与排名

积分 - 5935

排名 - 17402

最新评论 XML

1. Re: [全能系统监控工具dstat](#)

感觉好高级的样子，我也下载来玩完

--花瓣奶牛

2. Re: [【原】Ubuntu13.04安装、卸载Gnome3.8](#)

马上应该有13.10了。

--杨琼

3. Re: [scala实现kmeans算法](#)

在oschina上一位大牛给我的指点，原文贴上，供跟多的孩纸学习：oldpig 发表于 2013-09-03 10:45 1. Source.getLinesr 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Array.fill 3. 如果你要测算法的计算时间，应将两条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 \(Java虚拟机\) 优化大全和案例实战\(175\)](#)

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\(132\)](#)

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 sar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\(1\)](#)

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