

光于前裕于后的博客

记录一下新学知识，如有侵权请告知

原 SparkR初体验

置顶 2016年05月11日 20:41:19

阅读量: 1

突然有个想法，R只能处理百万级别的数据，如果R能运行在spark上多好！搜了下发现13年SparkR这个项目就启动了，感谢美帝！

- 1.你肯定得先装个spark吧。看这：[Spark本地模式与Spark Standalone伪分布模式](#)
- 2.你肯定得会R吧。看这：[R语言入门](#)
- 3.启动SparkR就可以了
- 3.1启动于本地（单机）

Spark also provides an experimental R API since 1.4 (only DataFrames APIs included).To run Spark interactively in a R inter
usebin/sparkR:

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

[plain]

```
1. guo@drquo:/opt/spark-1.6.1-bin-hadoop2.6$ ./bin/sparkR #这样直接运行默认在本地运行，相当于sparkR --master local[2]
2. R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
3. Copyright (C) 2015 The R Foundation for Statistical Computing
4. Platform: x86_64-pc-linux-gnu (64-bit)
5.
6. R是自由软件，不带任何担保。
7. 在某些条件下你可以将其自由散布。
8. 用'license()'或'licence()'来看散布的详细条件。
9.
10. R是个合作计划，有许多人为之做出了贡献。
11. 用'contributors()'来看合作者的详细情况
12. 用'citation()'会告诉你如何在出版物中正确地引用R或R程序包。
13.
14. 用'demo()'来看一些示范程序，用'help()'来阅读在线帮助文件，或
15. 用'help.start()'通过HTML浏览器来看帮助文件。
16. 用'q()'退出R。
17.
18. Launching java with spark-submit command /opt/spark-1.6.1-bin-hadoop2.6/bin/spark-submit "sparkr-shell" /tmp/RtmpmkEgRV/backend_port215
83a90cfc4
19. 16/05/12 03:30:35 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appli
cable
20.
21. Welcome to
22.
23.  _/ _/ _/ _/ _/ _/ _/
24. _\ V _ V _ _/ _/ _/
25. / _/ . _/ \ _/ _/ / _/ \ _/ version 1.6.1
26.  / _/
27.
28.
29. Spark context is available as sc, SQL context is available as sqlContext
```

3.2启动于Spark Standalone集群，别忘了先启动集群。

[plain]

```
1. guo@drquo:/opt/spark-1.6.1-bin-hadoop2.6$ bin/sparkR --master spark://drquo:7077
2.
3. Launching java with spark-submit command /opt/spark-1.6.1-bin-hadoop2.6/bin/spark-submit "--master" "spark://drquo:7077" "sparkr-shel
l" /tmp/RtmpXmU51Q/backend_port23516636af0a
4. 16/05/12 11:08:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appli
cable
5.
```

联系我们



关于 招
©2018 CSD
百度提供

经营性网站
网络110报警
中国互联网
北京互联网

3.3启动于yarn, 别忘了先启动yarn和hdfs

联系我们



关于 招聘
©2018 CSD
百度提供

经营性网站备案
网络110报警
中国互联网举
北京互联网通

```

[plain]
1. #读入本地数据框
2. > localDF <- data.frame(name=c("John", "Smith", "Sarah"), age=c(19, 23, 18))
3. > localDF
4.      name age
5. 1 John   19
6. 2 Smith  23
7. 3 Sarah  18
8. > df <- createDataFrame(sqlContext, localDF)
9. > printSchema(df)
10. root
11. |-- name: string (nullable = true)
12. |-- age: double (nullable = true)
13. #从本地文件读入
14. > peopleDF<-read.df(sqlContext,"people.json","json")
15. > peopleDF
16. DataFrame[age:bigint, name:string]
17. > head(peopleDF)
18.      age  name
19. 1 NA Michael
20. 2 30   Andy
21. 3 19   Justin
22. > peopleC <- collect(peopleDF)
23. > print(peopleC)
24.      age  name
25. 1 NA Michael
26. 2 30   Andy
27. 3 19   Justin
28. > printSchema(peopleDF)
29. root
30. |-- age: long (nullable = true)
31. |-- name: string (nullable = true)
32. > registerTempTable(peopleDF, "people")
33. #执行sql语句
34. > teenagers <- sql(sqlContext, "SELECT name FROM people WHERE age >= 13 AND age <= 19")
35. > teenagersLocalDF <- collect(teenagers)
36. > head(teenagersLocalDF)
37.      name
38. 1 Justin
39. > teenagers
40. DataFrame[name:string]
41. > print(teenagersLocalDF)
42.      name
43. 1 Justin
44. #还可以用hive sql呢!
45. > hiveContext <- sparkRHive.init(sc)
46. 16/05/12 13:16:18 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
47. 16/05/12 13:16:18 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
48. 16/05/12 13:16:25 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0
49. 16/05/12 13:16:25 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException

```

```
50. 16/05/12 13:16:28 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
51. 16/05/12 13:16:29 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
52. > sql(hiveContext, "CREATE TABLE IF NOT EXISTS src (key INT, value STRING)")
53. DataFrame[result:string]
54. > sql(hiveContext, "LOAD DATA LOCAL INPATH 'examples/src/main/resources/kv1.txt' INTO TABLE src")
55. DataFrame[result:string]
56. > results <- sql(hiveContext, "FROM src SELECT key, value")
57. > head(results)
58.   key  value
59. 1 238 val_238
60. 2   86 val_86
61. 3 311 val_311
62. 4   27 val_27
63. 5 165 val_165
64. 6 409 val_409
65. > print(results)
66. DataFrame[key:int, value:string]
67. > print(collect(results))
68.   key  value
69. 1 238 val_238
70. 2   86 val_86
71. 3 311 val_311
```

更多操作请看官方文档：<https://spark.apache.org/docs/latest/sparkr.html>

看一下drguo:4040，有了八个已完成的job

← → ⓘ drguo:4040/jobs/ 🔍 搜索 ☆ 自 📄 ⬇️ 🏠 ⌂

Spark 1.6.1

Jobs Stages Storage Environment Executors SQL SQL1 SparkR applica

Spark Jobs (?)

Total Uptime: 26 min
Scheduling Mode: FIFO
Completed Jobs: 8

▶ Event Timeline

Completed Jobs (8)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:45:33	0.1 s	1/1	2/2
6	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:43:28	68 ms	1/1	1/1
5	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:34:43	0.1 s	1/1	2/2
4	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:31:06	62 ms	1/1	2/2
3	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:27:21	42 ms	1/1	1/1
2	dfToCols at NativeMethodAccessorImpl.java:-2	2016/05/14 13:27:21	0.4 s	1/1	1/1
1	loadDF at NativeMethodAccessorImpl.java:-2	2016/05/14 13:25:20	0.4 s	1/1	2/2
0	collectPartitions at NativeMethodAccessorImpl.java:-2	2016/05/14 13:21:41	0.4 s	1/1	1/1

再看一下最后一个job的详细信息

<http://blog.csdn.net/>

联系我们

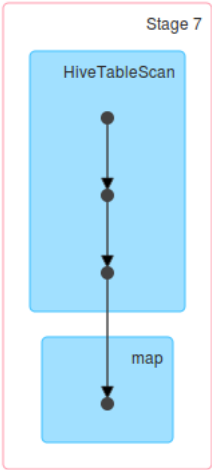


关于 招聘
©2018 CSDN
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

Completed Stages: 1

- ▶ Event Timeline
- ▼ DAG Visualization



Completed Stages (1)

Stage Id	Description
7	dfToCols at NativeMethodAccessorImpl.java:-2 http://blog.csdn.net/

```
[plain]
1. > getwd()
2. [1] "/opt/spark-1.6.1-bin-hadoop2.6"
3. > setwd("/home/guo/RWorkSpaces")
4. > getwd()
5. [1] "/home/guo/RWorkSpaces"
6. > x<-c(1,1,2,2,3,3,3)
7. > y<-c("女","男","女","男","女","男","女")
8. > z<-c(80,85,92,76,61,95,88)
9. > student<-data.frame(class=x,sex=y,score=z)
10. > student
11.      class sex score
12. 1      1  女   80
13. 2      1  男   85
14. 3      2  女   92
15. 4      2  男   76
16. 5      3  女   61
17. 6      3  男   95
18. 7      3  女   88
19. > row.names(student)<-c("凤姐","波多","杰伦","毕老爷","波","杰","毕老")#改变行名
20. > student
21.      class sex score
22. 凤姐      1  女   80
23. 波多      1  男   85
24. 杰伦      2  女   92
25. 毕老爷    2  男   76
26. 波        3  女   61
27. 杰        3  男   95
28. 毕老      3  女   88
29. > student$score
30. [1] 80 85 92 76 61 95 88
31. > student[,3]
32. [1] 80 85 92 76 61 95 88
33. > student[,score]
34. Error in `[.data.frame`(student, , score) : 找不到对象'score'
35. > student[, "score"]
36. [1] 80 85 92 76 61 95 88
37. > student[["score"]]
38. [1] 80 85 92 76 61 95 88
39. > student[[3]]
40. [1] 80 85 92 76 61 95 88
41. > student[1:2,1:3]
42.      class sex score
43. 凤姐      1  女   80
44. 波多      1  男   85
45. > student[student$score>80,]
46.      class sex score
47. 波多      1  男   85
48. 杰伦      2  女   92
```

联系我们



关于 招聘
©2018 CSD
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

```
49. 杰      3 男   95
50. 毕老   3 女   88
51. > attach(student)
52. > student[score>80,]
53.      class sex score
54. 波多     1 男   85
55. 杰伦     2 女   92
56. 杰      3 男   95
57. 毕老     3 女   88
```

5.提交R程序

```
[plain]
1. guo@drquo:/opt/spark-1.6.1-bin-hadoop2.6$ ./bin/spark-submit examples/src/main/r/dataframe.R
```

dataframe.R

```
[plain]
1. library(SparkR)
2.
3. # Initialize SparkContext and SQLContext
4. sc <- sparkR.init(appName="SparkR-DataFrame-example")
5. sqlContext <- sparkRSQL.init(sc)
6.
7. # Create a simple local data.frame
8. localDF <- data.frame(name=c("John", "Smith", "Sarah"), age=c(19, 23, 18))
9.
10. # Convert local data frame to a SparkR DataFrame
11. df <- createDataFrame(sqlContext, localDF)
12.
13. # Print its schema
14. printSchema(df)
15. # root
16. # |-- name: string (nullable = true)
17. # |-- age: double (nullable = true)
18.
19. # Create a DataFrame from a JSON file
20. path <- file.path(Sys.getenv("SPARK_HOME"), "examples/src/main/resources/people.json")
21. peopleDF <- read.json(sqlContext, path)
22. printSchema(peopleDF)
23.
24. # Register this DataFrame as a table.
25. registerTempTable(peopleDF, "people")
26.
27. # SQL statements can be run by using the sql methods provided by sqlContext
28. teenagers <- sql(sqlContext, "SELECT name FROM people WHERE age >= 13 AND age <= 19")
29.
30. # Call collect to get a local data.frame
31. teenagersLocalDF <- collect(teenagers)
32.
33. # Print the teenagers in our dataset
34. print(teenagersLocalDF)
35.
36. # Stop the SparkContext now
37. sparkR.stop()
```

官方文档：<https://spark.apache.org/docs/latest/api/R/index.html>

<https://spark.apache.org/docs/latest/sparkr.html>

下面转自：<http://mt.sohu.com/20151023/n424011438.shtml> 作者：孙锐，英特尔大数据团队工程师，HIVE和Shark项目贡献者，SparkR主力贡献者之一。

R和Spark的强强结合应运而生。2013年9月SparkR作为一个独立项目启动于加州大学伯克利分校的大名鼎鼎的AMPLAB实验室，与Spark源出同门。2014年1月，SparkR项目在github上开源（<https://github.com/amplab-extras/SparkR-pkg>）。随后，来自工业界的Alteryx、Databricks、Intel等公司和来自学术界的普渡大学，以及其它开发者积极参与到开发中来，最终在2015年4月成功地合并进Spark代码库的主干分支，并在Spark 1.4版本中作为重要的新特性之一正式宣布。

当前特性SparkR往Spark中增加了R语言API和运行时支持。Spark的API由Spark Core的API以及各个内置的高层组件（Spark Streaming, Spark SQL, ML Pipelines和MLlib, Graphx）的API组成，目前SparkR只提供了Spark的两组API的R语言封装，即Spark Core的RDD

联系我们



关于 招聘
©2018 CSDN
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

API和Spark SQL的DataFrame API。

需要指出的是，在Spark 1.4版本中，SparkR的RDD API被隐藏起来没有开放，主要是出于两点考虑：

RDD API虽然灵活，但比较底层，R用户可能更习惯于使用更高层的API；

RDD API的实现上目前不够健壮，可能会影响用户体验，比如每个分区的数据必须能全部装入到内存中的限制，对包含复杂数据的RDD的处理可能会存在问题等。

目前社区正在讨论是否开放RDD API的部分子集，以及如何在RDD API的基础上构建一个更符合R用户习惯的高层API。

RDD API用户使用SparkR RDD API在R中创建RDD，并在RDD上执行各种操作。

目前SparkR RDD实现了Scala RDD API中的大部分方法，可以满足大多数情况下的使用需求：

SparkR支持的创建RDD的方式有：

从R list或vector创建RDD (parallelize())

从文本文件创建RDD (textFile())

从object文件载入RDD (objectFile())

SparkR支持的RDD的操作有：

数据缓存，持久化控制：cache(),persist(),unpersist()

数据保存：saveAsTextFile(), saveAsObjectFile()

常用的数据转换操作，如map(),flatMap(),mapPartitions()等

数据分组、聚合操作，如partitionBy(),groupByKey(),reduceByKey()等

RDD间join操作，如join(), fullOuterJoin(), leftOuterJoin()等

排序操作,如sortBy(), sortByKey(), top()等

Zip操作，如zip(), zipWithIndex(), zipWithUniqueld()

重分区操作，如coalesce(), repartition()

其它杂项方法

和Scala RDD API相比，SparkR RDD API有一些适合R的特点：

SparkR RDD中存储的元素是R的数据类型。

SparkR RDD transformation操作应用的是R函数。

RDD是一组分布式存储的元素，而R是用list来表示一组元素的有序集合，因此SparkR将RDD整体上视为一个分布式的list。Scala API中RDD的每个分区的数据由iterator来表示和访问，而在SparkR RDD中，每个分区的数据用一个list来表示，应用到分区的转换操作，如mapPartitions()，接收到的分区数据是一个list而不是iterator。

为了符合R用户经常使用lapply()对一个list中的每一个元素应用某个指定的函数的习惯，SparkR在RDD类上提供了SparkR专有的transformation方法：lapply()、lapplyPartition()、lapplyPartitionsWithIndex()，分别对应于Scala API的map()、mapPartitions()、mapPartitionsWithIndex()。

DataFrame API Spark 1.3版本引入了DataFrame API。相较于RDD API，DataFrame API更受社区的推崇，这是因为：

DataFrame的执行过程由Catalyst优化器在内部进行智能的优化，比如过滤器下推，表达式直接生成字节码。

基于Spark SQL的外部数据源（external data sources）API访问（装载，保存）广泛的第三方数据源。

使用R或Python的DataFrame API能获得和Scala近乎相同的性能。而使用R或Python的RDD API的性能比起Scala RDD API来有较大的性能差距。

Spark的DataFrame API是从R的 Data Frame数据类型和Python的pandas库借鉴而来，因而对于R用户而言，SparkR的DataFrame API是很自然的。更重要的是，SparkR DataFrame API性能和Scala DataFrame API几乎相同，所以推荐尽量用SparkR DataFrame来编程。

目前SparkR的DataFrame API已经比较完善，支持的创建DataFrame的方式有：

从R原生data.frame和list创建

联系我们



关于 招聘
©2018 CSDN
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

从SparkR RDD创建

从特定的数据源(JSON和Parquet格式的文件)创建

从通用的数据源创建

将指定位置的数据源保存为外部SQL表，并返回相应的DataFrame

从Spark SQL表创建

从一个SQL查询的结果创建

支持的主要的DataFrame操作有：

·数据缓存，持久化控制：cache(),persist(),unpersist()

数据保存：saveAsParquetFile(), saveDF()（将DataFrame的内容保存到一个数据源），saveAsTable()（将DataFrame的内容保存为数据源的一张表）

集合运算：unionAll(), intersect(), except()

Join操作：join(), 支持inner、full outer、left/right outer和semi join。

数据过滤：filter(), where()

排序：sortDF(), orderBy()

列操作：增加列– withColumn(), 列名更改– withColumnRenamed(), 选择若干列 –select()、selectExpr()。为了更符合R用户习惯，SparkR还支持用\$、[]、[[]操作符选择列，可以用\$<列名> <- 的语法来增加、修改和删除列

RDD map类操作：lapply()/map(), flatMap(), lapplyPartition()/mapPartitions(), foreach(), foreachPartition()

数据聚合：groupBy(), agg()

转换为RDD：toRDD(), toJSON()

转换为表：registerTempTable(),insertInto()

取部分数据：limit(), take(), first(), head()

编程示例总体上看，SparkR程序和Spark程序结构很相似。

基于RDD API的示例

要基于RDD API编写SparkR程序，首先调用sparkR.init()函数来创建SparkContext。然后用SparkContext作为参数，调用parallelize()或者textFile()来创建RDD。有了RDD对象之后，就可以对它们进行各种transformation和action操作。下面的代码是用SparkR编写的Word Count示例：

```
library(SparkR) #初始化SparkContext sc <- sparkR.init("local", "RWordCount") #从HDFS上的一个文本文件创建RDD lines <- textFile(sc, "hdfs://localhost:9000/my_text_file") #调用RDD的transformation和action方法来计算word count #transformation用的函数是R代码 words <- flatMap(lines, function(line) { strsplit(line, " ")[[1]] }) wordCount <- lapply(words, function(word) { list(word, 1L) }) counts <- reduceByKey(wordCount, "+", 2L) output <- collect(counts)
```

基于DataFrame API的示例

基于DataFrame API的SparkR程序首先创建SparkContext，然后创建SQLContext，用SQLContext来创建DataFrame，再操作DataFrame里的数据。下面是用SparkR DataFrame API计算平均年龄的示例：library(SparkR) #初始化SparkContext和SQLContext sc <- sparkR.init("local", "AverageAge") sqlCtx <- sparkRSQL.init(sc) #从当前目录的一个JSON文件创建DataFrame df <- jsonFile(sqlCtx, "persons.json") #调用DataFrame的操作来计算平均年龄 df2 <- agg(df, age="avg") averageAge <- collect(df2)[1, 1]

对于上面两个示例要注意的一点是SparkR RDD和DataFrame API的调用形式和Java/Scala API有些不同。假设rdd为一个RDD对象，在Java/Scala API中，调用rdd的map()方法的形式为：rdd.map(...)，而在SparkR中，调用的形式为：map(rdd, ...)。这是因为SparkR使用了R的S4对象系统来实现RDD和DataFrame类。

架构SparkR主要由两部分组成：SparkR包和JVM后端。SparkR包是一个R扩展包，安装到R中之后，在R的运行时环境里提供了RDD和DataFrame API。

联系我们



关于 招聘
©2018 CSDN
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

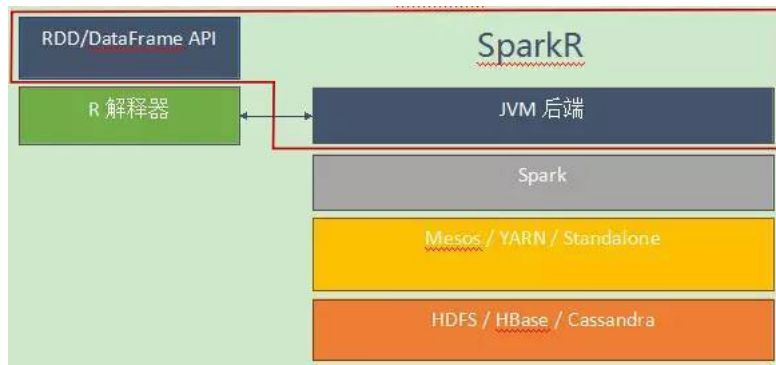


图1 SparkR软件栈

SparkR的整体架构如图2所示。

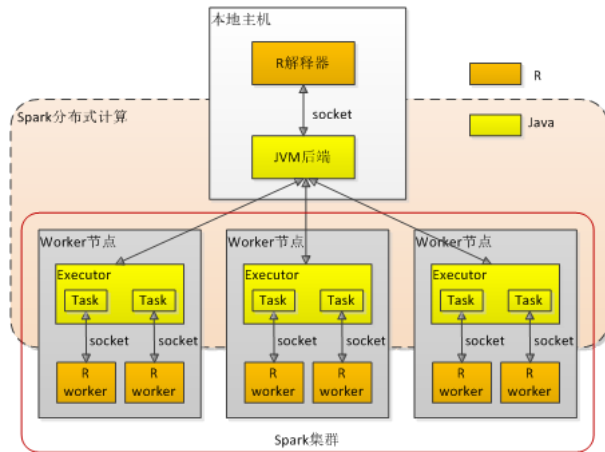


图2 SparkR架构

R JVM后端SparkR API运行在R解释器中，而Spark Core运行在JVM中，因此必须有一种机制能让SparkR API调用Spark Core的服务。R JVM后端是Spark Core中的一个组件，提供了R解释器和JVM虚拟机之间的桥接功能，能够让R代码创建Java类的实例、调用Java对象的实例方法或者Java类的静态方法。JVM后端基于Netty实现，和R解释器之间用TCP socket连接，用自定义的简单高效的二进制协议通信。

R Worker

SparkR RDD API和Scala RDD API相比有两大不同：SparkR RDD是R对象的分布式数据集，SparkR RDD transformation操作应用的是R函数。SparkR RDD API的执行依赖于Spark Core但运行在JVM上的Spark Core既无法识别R对象的类型和格式，又不能执行R的函数，因此如何在Spark的分布式计算核心的基础上实现SparkR RDD API是SparkR架构设计的关键。

SparkR设计了Scala RRDD类，除了从数据源创建的SparkR RDD外，每个SparkR RDD对象概念上在JVM端有一个对应的RRDD对象。RRDD派生自RDD类，改写了RDD的compute()方法，在执行时会启动一个R worker进程，通过socket连接将父RDD的分区数据、序列化后的R函数以及其它信息传给R worker进程。R worker进程反序列化接收到的分区数据和R函数，将R函数应用到分区数据上，再把结果数据序列化成字节数组传回JVM端。

从这里可以看出，与Scala RDD API相比，SparkR RDD API的实现多了几项开销：启动R worker进程，将分区数据传给R worker和R worker将结果返回，分区数据的序列化和反序列化。这也是Spark RDD API相比Scala RDD API有较大性能差距的原因。

DataFrame API的实现

由于SparkR DataFrame API不需要传入R语言的函数（UDF()方法和RDD相关方法除外），而且DataFrame中的数据全部是以JVM的数据类型存储，所以和SparkR RDD API的实现相比，SparkR DataFrame API的实现简单很多。R端的DataFrame对象就是对应的JVM端DataFrame对象的wrapper，一个DataFrame方法的实现基本上就是简单地调用JVM端DataFrame的相应方法。这种情况下，R Worker就不需要了。这是使用SparkR DataFrame API能获得和Scala API近乎相同的性能的原因。

当然，DataFrame API还包含了一些RDD API，这些RDD API方法的实现是先将DataFrame转换成RDD，然后调用RDD的相关方法。

展望SparkR目前来说还不是非常成熟，一方面RDD API在对复杂的R数据类型的支持、稳定性和性能方面还有较大的提升空间，另一方面DataFrame API在功能完备性上还有一些缺失，比如对用R代码编写UDF的支持、序列化/反序列化对嵌套类型的支持，这些问题相信会在后续的开发中得到改善和解决。如何让DataFrame API对熟悉R原生Data Frame和流行的R package如dplyr的用户更友好是一个有意思的方向。此外，下一步的开发计划包含几个大的特性，比如普渡大学正在做的在SparkR中支持Spark Streaming，还有Databricks正在做的在

联系我们



关于 招聘
©2018 CSD
百度提供



腾冲六日



SparkR中支持ML pipeline等。SparkR已经成为Spark的一部分，相信社区中会有越来越多的人关注并使用SparkR，也会有更多的开发与对SparkR的贡献，其功能和使用性将会越来越强。

总结Spark将正式支持R API对熟悉R语言的数据科学家是一个福音，他们可以在R中无缝地使用RDD和Data Frame API，借助Spark内存计算、统一软件栈上支持多种计算模型的优势，高效地进行分布式数据计算和分析，解决大规模数据集带来的挑战。工欲善其事，必先利其器，SparkR必将成为数据科学家在大数据时代的又一门新利器。

版权声明：本文为博主原创文章，转载请加上原文地址，谢谢！ https://blog.csdn.net/Dr_Guo/article/details/51377784

个人分类：[Spark](#) [R](#) [大数据动物园](#)

查看更多>>

想对作者说点什么？

我来说两句



IT影风 2017-07-07 19:09:26 #4楼

sparkR读取Mysql的一张表是不是把整张表加载到内存中的？这样的话数据不是很多了吗？最近在spark跑一个7000万条的数据,发现比单机的数据还慢,你觉得这样怎么解析？



qsdcr 2016-09-30 10:47:58 #3楼

df=createDataFrame(sqlContext,faithful) head(df) 楼主在执行head的时候没遇到问题么？



昵称字符数 2016-05-13 18:00:26 #2楼

码住 偏偏集群上spark1.4没有R要哭死了

[查看回复\(2\)](#)

[查看 7 条热评](#)

利用R语言实现spark大数据分析与可视化

系统概述 在日常业务分析中，R是非常常用的分析工具，而当数据量较大时，用R语言需要需用更多的时间来完成训练模型，spark作为大规模数据计算框架，采用内存计算，可以短时间内完成大量...



LW_GHY 2017-02-23 21:20:14 阅读数：2782

sparklyr包：实现Spark与R的接口+sparklyr 0.5

本文转载于雪晴数据网 日前，Rstudio公司发布了sparklyr包。该包具有以下几个功能：实现R与Spark的连接—sparklyr包提供了一个完整的dplyr后端筛选并聚合Spark...



sinat_26917383 2016-10-07 13:24:00 阅读数：4325

利用R语言实现spark大数据分析与可视化 - CSDN博客

系统概述 在日常业务分析中，R是非常常用的分析工具,而当数据量较大时,用R语言需要需用更多的时间来完成训练模型,spark作为大规模数据计算框架,采用内存计算,可以...

2018-4-24

Spark组件之SparkR学习5--R语言函数调用(跨文件调用) - CSDN博客

环境: RStudio R-3.2.1 Spark组件之SparkR学习5--R语言函数调用(跨文件调用) 1.在文件夹func下新建R文件addTest.R: 文件路径:D:/all/R/RStudio/R...

2018-6-6

早知道痔疮这么简单就能好，还做什么手术~

黄河医院 · 顶新

联系我们



关于 招聘

©2018 CSDN

百度提供

经营性网站备案

网络110报警

中国互联网

北京互联网

Hadoop+Spark+R+SparkR集群环境搭建

2017年11月15日 230KB 下载

Spark组件之SparkR学习4--Eclipse下R语言环境搭建 - CSDN博客

[1] "SparkR" 16/04/20 15:20:43 INFO SparkContext: Starting job: collectPartitions at NativeMethodAccessorImpl.java:-2 16/04/20 15:20:43 INFO ...
2018-6-1

Hadoop+Spark+R+SparkR集群环境搭建

Hadoop Spark R SparkR 大数据集群 安装文档。全是原生组件,部署在Centos系统上... Hadoop Spark R SparkR 大数据集群 ...SparkR初体验 利用R语言ark大...
2018-5-4

sparkR-入门知识

一、sparkR的简介 SparkR是一个R语言包，它提供了轻量级的方式使得可以在R语言中使用Apache Spark。在Spark 1.4中，SparkR实现了分布式的数据e，支持类...
qq_34941023 2016-07-09 18:55:29 阅读数：2777

SparkR

SparkR提供了轻量级的方式在R中使用Spark,SparkR实现了分布式的数据frame,支持类似查询，过滤和聚合等，(类似R中data frames : dplyr)，这个可以规模的数...
Yaphat 2016-11-10 16:31:22 阅读数：1655

大数据工具比较:R语言和Spark谁更胜一筹? - jarth的专栏 - 博客...

摘要:本文有两重目的,一是在性能方面快速对比下R语言和Spark,二是想向大家介绍下 Spark的机器学习库。背景 介绍 由于R语言本身是单线程的,所以可能从性能方面...
2017-1-19

讲解Spark API 最好的资料 - CSDN博客

作为主要开发语言,同时为了方便更多语言背景的人使用,还支持Java、Python和R语言...举报内容: 讲解Spark API 最好的资料 举报原因: 色情 政治 抄袭 广告 招聘 ...
2018-5-29

SparkR初探

这样看来，大部分R的分析，都能够直接跑在spark集群上了，再联想到去年Esri发布了ArcGIS对R语言的支持，可以预料到不远的未来，所有的集群运算都将被融为一体。...
allenlu2008 2016-05-21 21:41:00 阅读数：1523

闺房秘闻：1分钟就完了？教你1招坚挺

名门府祗 · 顶新

sparklyr包:实现Spark与R的接口+sparklyr 0.5 - CSDN博客

本文转载于雪晴数据网 日前,Rstudio公司发布了sparklyr包。该包具有以下几个功能: 实现R与Spark的连接—spa...
2018-5-21

R or Spark - CSDN博客

Spark or R 前天下班浏览朋友圈,雪晴数据网转发了一篇译文,大数据工具比较:R 语言和 Spark 谁更胜一筹?,原作者测试了在限定为单机环境下,使用Kaggle提供的手写...
2018-1-27

SparkR数据分析

本文的运行环境是ubuntu,在阅读这篇文章前,请先保证你已经成功配置了Spark,并设置好了全局变量 SPARK_HOME以及 PATH ,能够成功运行Spark.(如果你在终端输入sparkR ...

联系我们



关于 招聘
©2018 CSD
百度提供:

经营性网站
网络110报警
中国互联网
北京互联网

 a358463121 2016-01-20 16:23:58 阅读数：1831

SparkR终极解决方案

原文地址：<http://blog.csdn.net/wangjunji34478/article/details/70906537> 问题： Spark支持sparkR需要安装R ...

 jiabiao1602 2017-08-14 00:15:43 阅读数：262

SparkR初体验 - CSDN博客

原文地址http://blog.csdn.net/dr_guo/article/details/51377784 SparkR初体验 2016年05月11日 20:41:19 13072 突然有个想法,R只能处理百万级别的数据, ...

2018-5-22

undefined

sparkR的一个运行的例子

在sparkR在配置完成的基础上, 本例采用spark on yarn模式, 介绍sparkR运行的一个例子。 在spark的安装目录下, /examples/src/main/r, 有一个da

 zhoudetiankong 2016-06-16 14:13:24 阅读数：1975


SparkR (R on Spark)

<http://spark.apache.org/docs/latest/sparkr.html> SparkR (R on Spark) OverviewSparkDataFra...

 u014032673 2017-01-10 17:31:11 阅读数：898

安装SparkR

必须条件: 1:安装好JDK 2:安装好R 步骤1: 运行R Shell [jifeng@feng03 R-3.1.1]\$ R R version 3.1.1 (2014-07-10) -- "S...

 wind520 2015-09-30 00:20:04 阅读数：6263

Python与R的争锋：大数据初学者该怎样选？

在当下, 人工智能的浪潮席卷而来。从AlphaGo、无人驾驶技术、人脸识别、语音对话, 到商城推荐系统, 金融业的风控, 量化运营、用户洞察、企业征信、智能投顾等, 人工智能的应用广泛渗透到各行各业, 也让数据科...

 seeyousoonhhh 2017-11-28 16:04:08 阅读数：596

Spark组件之SparkR学习3--使用spark-submit向集群提交R代码文件data-manipulation.R

1.数据准备: 1.1 下载数据文件 wget <http://s3-us-west-2.amazonaws.com/sparkr-data/flights.csv> 1.2 上传到hdfs: hado...

 bob601450868 2016-04-20 13:00:07 阅读数：2703

程序员不会英语怎么行？

老司机教你一个数学公式秒懂天下英语



基于spark1.4.1的sparkR的实例操作

原文地址：<http://blog.csdn.net/bdchome/article/details/48104537> [Author]: kwu 基于spark1.4.1的spark...

 jiabiao1602 2017-08-24 18:17:27 阅读数：197

SparkR 1.4.0 的安装及使用

1、./sparkR打开R shell之后, 使用不了SparkR的函数 [root@master sparkR]#./bin/sparkR 能进入R, 和没装SparkR的一样, 无报错 > libra...

 wa2003 2015-06-25 13:27:04 阅读数：2369

在R或Rstudio中调用SparkR

libpath libpath .libPaths(libpath) rm(libpath) library(rJava) library(devtools) library(Sp...

 u010022051 2016-04-11 09:26:43 阅读数：3388

数据科学家如何优雅的运行R在spark内存计算引擎上

来源:<http://www.ppvke.com/Blog/archives/46156> R在数据科学中超过10,000包, 是主要的编程语言之一。R是开源软件, 作为 统计学和计算机科学课程的一...

 jiabiao1602 2017-08-14 00:13:54 阅读数：399

联系我们




关于 招聘
©2018 CSD
百度提供:

经营性网站
网络110报警
中国互联网
北京互联网

Spark入门基础教程

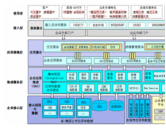
Spark入门基础教程

 lbyyy 2016-11-25 14:37:03 阅读数：15342

云计算架构图

云计算的体系结构

百度广告



R | sparkR的安装与使用、函数尝试笔记、一些案例

本节内容转载于博客： wa2003 —————一、SparkR 1.4.0 的安装及使用1、R打开R shell之后，使用...

 sinat_26917383 2016-12-01 15:14:03 阅读数：1482


Spark学习——利用Scala语言开发Spark应用程序

Spark学习——利用Scala语言开发Spark应用程序Spark内核是由Scala语言开发的，因此使用Scala语言开发Spark应用程序是自然而然的事情。如果你对语言还不太熟悉，可以阅...

 sbq63683210 2016-06-07 20:59:21 阅读数：4141

Spark入门-什么是Spark

·spark认识 Spark使用Scala语言进行实现，它是一种面向对象、函数式编程语言，能够像操作本地集合对象一样轻松地操作分布式数据集，在Spark官网介绍，它具有运行速度快、易用性好、通用性...

 u014372225 2016-07-16 21:13:40 阅读数：2567

Apache Spark 2.2.0 中文文档 - SparkR (R on Spark) | ApacheCN

SparkR (R on Spark) 概述 SparkDataFrame 启动: SparkSession 从 RStudio 来启动 创建 SparkDataFrames...

 u010859707 2017-09-26 12:40:25 阅读数：495

Spark核心RDD：Sort排序详解

1.sortByKey 无可非议sortByKey是Spark的最常用的排序，简单的案例暂且跳过，下面给一个非简单的案例，让我进入排序之旅 对下面简单元祖，要求先按元素1升序，若元素1相同，则按元素3...

 jiangpeng59 2016-10-26 23:27:28 阅读数：13706

C#程序加壳，虚拟机外壳，强度堪比VMP

集自动代码移植、混淆、外壳加密于一身，无需编程就能达到极高的保护强度




深入了解spark运行计划及调优

问题导读 1.首次运行hive-console需要什么条件？ 2.运行hive/console是否需要启动Spark？ 3.如何查看查询的Unresolved LogicalPlan？ 4...

 javastart 2016-02-07 16:30:20 阅读数：1706

DataFrame registerTempTable(注册临时表)后Table Not Found问题的解决

转：http://blog.csdn.net/sparkexpert/article/details/51206487 所说这个错误没有遇到过不过背后的原理还是需要知道的。将数据存成...

 weixin_36630761 2017-10-16 11:01:02 阅读数：68

spark-1.6.x-总结

spark-1.6.0-总结

 high2011 2016-08-07 18:08:42 阅读数：1141

WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-

[root@db96 hadoop]# hadoop dfs -put ./in DEPRECATED: Use of this script to execute hdfs command is ...

 wulantian 2014-07-25 11:48:08 阅读数：34324

联系我们



关于 招聘

©2018 CSD

百度提供

经营性网站

网络110报警

中国互联网

北京互联网

当对象包含嵌套对象时，使用Spark SQL执行sql查询抛出scala.MatchError异常

1. 运行环境 本文使用Spark SQL 2.1.0版本 2. 使用代码 例如有个对象，除了包含简单的基本数据String，int之外还包含一个Location对象，就是所说的嵌

gangchengzhong 2017-04-13 09:44:49 阅读数：2699

呼叫中心系统

登录呼叫中心管理系统

百度广告



SparkSQL学习笔记(二)DataSource

简介： SparkSQL通过DataFrame接口支持处理各种数据源，df可以抽象为RDD或注册内存(临时)表处理，临时表可以通过SQL操作并返回一个结果df。
ave 方法： 最简单...

wcdc0521 2015-11-24 13:18:50 阅读数：7244

Spark DataFrames入门指南：创建和操作DataFrame

一、从csv文件创建DataFrame 本文将介绍如何从csv文件创建DataFrame。如何做？ 从csv文件创建DataFrame主要包括以下步骤： 1、
d....

LW_GHY 2016-05-23 11:56:00 阅读数：43812

Spark 2.0系列之SparkSession详解

用Spark的各项功能，用户不但可以使用DataFrame和Dataset的各种API，学习Spark的难度也会大大降低。本文就SparkSession在Spark2 0中的功能和以阐释。 ...

u013063153 2017-01-19 13:50:14 阅读数：14453

Spark之SparkSession

最近学习SparkSql时接触了SparkSession。SparkSession是Spark 2.0引如的新概念。SparkSession为用户提供了统一的切入点，来让用户学习spark的各项功能。 ...

u012430664 2017-02-27 13:04:52 阅读数：4123

spark零基础学习路线指导

spark零基础学习路线指导 问题导读 1.你认为spark该如何入门？ 2.你认为spark入门编程需要哪些步骤？ 3.本文介绍了spark哪些编程知识？ s...

SCGH_Fx 2017-07-04 13:56:38 阅读数：6268

码农怎能不懂英语？！试试这个数学公式

老司机教你一个数学公式秒懂天下英语



spark入门详解

1. Spark中的基本概念 在Spark中，有下面的基本概念。Application:基于Spark的用户程序，包含了一个driver program和集群中多个executor Dri...

anningzhu 2017-03-07 20:24:36 阅读数：4394

Spark学习笔记--Spark基础知识

1、RDD表示分布在多个计算节点上的可以并行操作的元素集合，是spark主要的变成抽象。Spark Streaming 是 Spark 提供的对实时数据进行流式计算的组件
Spark是一个用于集...

a1628864705 2016-09-08 22:57:56 阅读数：2435

Spark入门三部曲之第一步Spark基础知识

Spark运行环境 Spark 是Scala写的, 运行在JVM上。所以运行环境是Java6或者以上。如果想要使用 Python API，需要安装Python 解释器2.6版本或者以上。
目前...

maixia24 2015-08-04 13:18:00 阅读数：9389

Spark基础入门（一）-----RDD基础

（一）、RDD定义 不可变 分布式对象集合 创建RDD有两种方式： （二）、RDD分区数 （三）、RDD操作 转换操作与行动操作...

联系我们



关于 招聘
©2018 CSD
百度提供:

经营性网站
网络110报警
中国互联网
北京互联网

 silviakafka 2017-01-12 15:29:03 阅读数：2605

Spark 入门实战之最好的实例

转载：<https://www.ibm.com/developerworks/cn/opensource/os-cn-spark-practice1/> 搭建开发环境 安装 Scala I...

 gongpulin 2016-05-29 23:27:28 阅读数：30465

TIOBE 编程语言排行榜

编程语言

百度广告



子雨大数据之Spark入门教程---Spark入门： Spark运行架构1.2

本节首先介绍Spark的基本概念和架构设计方法，然后介绍Spark运行基本流程。基本概念 在具体讲解Spark运行架构之前，需要先了解几个重要的概念：RDD：是弹性分布式数据集（Re...

 u011630575 2017-02-23 00:33:42 阅读数：1072

Spark快速入门指南

- Spark是什么？ Spark is a MapReduce-like cluster computing framework designed to support low-latency...

 macyang 2011-12-24 23:34:48 阅读数：19893

Spark 学习入门教程

转载请注明作者，谢谢支持！ 一、环境准备 测试环境使用的cdh提供的quickstart vm hadoop版本： 2.5.0-cdh5.2.0 spark版本： 1.1.0 二、H...

 wankunde 2014-12-02 10:12:51 阅读数：94059

spark入门介绍(菜鸟必看)

什么是Spark Apache Spark是一个围绕速度、易用性和复杂分析构建的大数据处理框架。最初在2009年由加州大学伯克利分校的AMPLab开发，并于2010年成为Apache的开源项目之...

 u011497897 2017-05-09 11:06:16 阅读数：559

Rstudio黑科技Sparklyr于Ubuntu系统的部署安装与调试

配置镜像源deb <https://bin/linux/ubuntu/zesty/> or deb <https://bin/linux/ubuntu/yakkety/> or d...

 Hello_Word____ 2017-07-10 15:13:03 阅读数：492

程序员不会英语怎么行？

老司机教你一个数学公式秒懂天下英语



使用Spark DataFrame进行大数据处理

简介 DataFrame让Spark具备了处理大规模结构化数据的能力，在比原有的RDD转化方式易用的前提下，计算性能更还快了两倍。这一个小小的API，隐含着Spark希望大一统「大数据江...

 vfgbv 2016-06-03 13:55:33 阅读数：4293

Ubuntu 下安装sparklyr 并连接远程spark集群

安装sparklyr1.通过devtools包实现sparklyr包的安装：install.packages("devtools") devtools::install_github("rstudio...

 The_One_is_all 2017-07-18 15:42:53 阅读数：918

Spark入门教程（1）——spark是什么及发展趋势概述

本教程源于2016年3月出版书籍《Spark原理、机制及应用》，如有兴趣，请支持正版书籍。随着互联网为代表的信息技术深度发展，其背后由于历史积累产生了TB、PB甚至EB级数据量，由于传统机器的软硬件...

 xwc35047 2016-04-06 09:41:30 阅读数：16628

Spark修炼之道（进阶篇）——Spark入门到精通：第七节 Spark运行原理

本节主要内容 Spark运行方式 Spark运行原理解析 本节内容及部分图片来自：[http://blog.csdn.net/book_mmicky/article/details/25714419...](http://blog.csdn.net/book_mmicky/article/details/25714419)

联系我们



关于 招聘
©2018 CSD
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

 lovehuangjiaju 2015-09-22 19:54:01 阅读数：16852

Spark入门

spark历史：伯克利实验室研究项目，基于Hadoop的Mapreduce机制，引入内存管理机制，提高了迭代式计算和交互式中的效率。 spark组件： spark co
k基本功能，包括任...

 sinat_32873711 2017-12-02 17:18:33 阅读数：170

没有更多推荐了， [返回首页](#)

个人资料



光于前裕于后

原创 粉丝 喜欢
106 **188** **190**

等级： **博客 5** 访问： 54万+
积分： 5298 排名： 6694
勋章： 

联系我们



关于 招聘
©2018 CSDN
百度提供

经营性网站备案
网络110报警
中国互联网
北京互联网

xps 13



最新文章

- Scala基础
- 使用Keras实现多层前馈神经网络对Iris（鸢尾花卉）数据集进行多分类
- 使用python获取pdf上的文字(in win10)
- Json格式字符串转Java对象
- MyBatis更新数据（输入参数类型为Map）

个人分类

深度学习与神经网络	1篇
大数据动物园	54篇
数据挖掘	16篇
Hadoop	40篇
Mahout	3篇

[展开](#)

归档

2018年5月	1篇
2018年4月	1篇
2018年3月	1篇
2018年1月	2篇
2017年11月	2篇

展开

热门文章

log4j.properties配置详解与实例
阅读量：64186

验证码识别（Tess4J初体验）
阅读量：31669

用R进行多元线性回归分析建模
阅读量：21876

MapReduce Input Split（输入分/切解）
阅读量：17777

SparkR初体验
阅读量：17005

最新评论

使用Ambari给HDP集群安装K...
Dr_Guo：[code=plain] addprinc -randkey min xst -no...

Hadoop HA高可用集群搭建（...
jie_linux：想问一下,公司要做HA 但是生...
用。如果格式化那数据不就没有了。如...
不格式化能否...

Mahout开发中发现缺少MySQL...
asibity：多谢，解决了问题

Scala基础
Dr_Guo：scala中看到参数类型有 => 表示此参...
数是函数，如(Int, Int) =&am...

log4j.properties配...
u012031380：收藏了，谢群主分享！

联系我们



关于 招聘

©2018 CSD

 百度提供:

经营性网站在

网络110报警

中国互联网举

北京互联网过