aggregate vs treeAggregate

原创 2016年08月04日 17:10:53 标签：spark 1614

aggregate

```
aggregate[U: ClassTag](zeroValue: U)(seqOp: (U, T) => U, combOp: (U, U) => U)
```

aggregate函数将每个分区进行seqOp,且从zeroValue开始遍历分区里的所有元素.然后用combOp,从zero Value开始遍历所有分区的结果.

注意:每个partition的seqOp只应用一次zeroValue,最后的combOp也应用一次zeroValue.

例子:

```
scala> def seq(a:Int,b:Int):Int={
     | println("seq:"+a+":"+b)
     | math.min(a,b)}
seq: (a: Int, b: Int)Int

scala> def comb(a:Int,b:Int):Int={
     | println("comb:"+a+":"+b)
     | a+b}
comb: (a: Int, b: Int)Int

val z =sc.parallelize(List(1,2,4,5,8,9),3)
scala> z.aggregate(3)(seq,comb)
seq:3:4
seq:3:1
seq:1:2
seq:3:8
seq:3:5
seq:3:9
comb:3:1
comb:4:3
comb:7:3
res10: Int = 10
```

treeAggregate

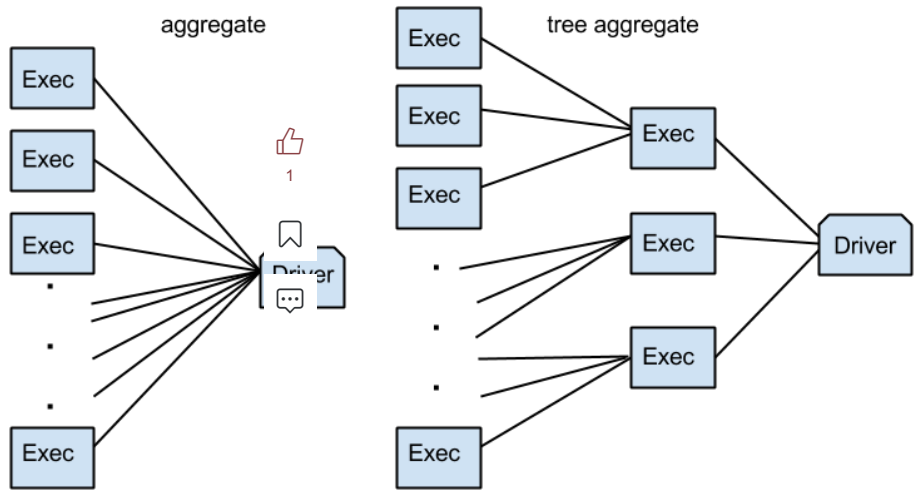
```
treeAggregate[U: ClassTag](zeroValue: U)(
      seqOp: (U, T) => U,
      combOp: (U, U) => U,
      depth: Int = 2)
```

与aggregate不同的地方是:在每个分区,会做两次或者多次combOp,避免将所有局部的值传给driver端.另外,经过测验初始值zeroValue不会参与combOp.

例子:

```
scala> z.treeAggregate(3)(seq,comb)
seq:3:4
seq:3:5
seq:3:1
seq:1:2
seq:3:8
seq:3:9
comb:3:3
comb:6:1
```

对比图:



注释:
Aggregate

1. each executor holds a portion of learning set
2. broadcast model to excutors
3. collect results to driver

TreeAggregate

1. simple heuristic to add level
2. perform partial aggregation by shipping results to other executors(by repartitioning)

针对Spark的RDD，API中有一个aggregate函数，本人理解起来费了很大劲，明白之后，mark一下，供以后参考。 首先，Spark文档中aggregate函数定义如下 def aggreg...

《利用python进行数据分析》学习笔记（一）

mike112223 2017年06月12日 19:08 436

处理usa.gov数据 导入数据 import json path = 'usagov_bitly_data2012-03-16-1331923249.txt' records = [json.l...

R: 矩阵运算及常用函数 II - aggregate

u013889309 2014年11月26日 17:05 2109

aggregate也是跟SAC有关系的一个函数(stats包中)：先将对象分解为不同的组别(回忆一下split函数)，然后分个处理，最后合并显示。 具体地说，aggregate()函数将数据集(依...

Mongo 聚合框架-Aggregate(一)

u013940812 2017年08月27日 23:10 651

一 概念1、简介 使用聚合框架可以对集合中的文档进行变换和组合。可以用多个构件创建一个管道，用于对一连串的文档进行处理。构件有：筛选、投射、分组、排序、限制和跳过。 MongoDB的聚合管道...

spark-aggregate与treeAggregate的理解

u011724402 2018年01月14日 16:41 48

spark-mllib中许多算法用到了treeAggregate这个方法，使用该方法而不是aggregate方法能够提升算法的性能。比如mllib中的GaussianMixture模型可以提升20%的...

aggregate vs treeAggregate

lookqlp 2016年08月04日 17:16 1614

aggregate与treeAggregate对比

Spark 之RDD API大全

tangshiweibbs 2017年04月11日 20:01 183

package scala import org.apache.spark.{SparkConf, SparkContext} /** * Created by root on 17-4-11...

结合源码彻底讲解Aggregate vs treeAggregate

Aggregate本文主要是讲解两个常见的聚合操作：aggregate vs treeAggregate首先讲解aggregate，该函数的方法具体名称如下：def aggregate[U: Clas...

rlnLo2pNEfx9c 2018年01月14日 00:00 116

treeAggregate和Aggregate的区别

lishuan182 2017年07月24日 15:41 141

【aggregate】 scala> def seq(a:Int,b:Int):Int={ | println("seq:"+a+":"+b) ...

RDD api整理

zbf8441372 2015年04月07日 16:51 7128

RDD[T]Transformations rdd api 备注 persist/cache map(f: T => U) keyBy(f: T => K) 特殊的m...

讲解 Spark API 最好的资料

kebu12345678 2018年02月12日 16:43 75

http://homepage.cs.latrobe.edu.au/zhe/ZhenHeSparkRDDAPIExamples.htmlOur research group has a very st...

MongoDB中group() mapReduce() aggregate()之比较

对于SQL而言，如果从users表里查询每个team所有成员的number， huntzw 2013年03月13日 19:26 8204

Scala aggregate

power0405hf 2015年12月17日 21:35 2014

1.Spark函数讲解：aggregate 2.Example of the Scala aggregate function1.Spark函数讲解：aggregate函数原型：def aggreg...

一个数学公式教你秒懂天下英语

老司机教你一个数学公式秒懂天下

1

深入机器学习系列2-SVM

mkt_transwarp 2017年08月24日 09:44 288

Support Vector Machine 支持向量 一种机器学习算法。

spark rdd 详解

mlljava1111 2016年03月30日 15:10 878

转：http://homepage.cs.latrobe.edu.au/zhe/ZhenHeSparkRDDAPIExamples.htmlaggregateThe aggregate functio...

OpenStack之Region, Availability Zone和Host Aggregate的理解

OpenStack是Amazon AWS的开源实现，直白点就是山寨产品吧，

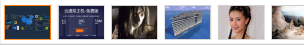
jiaomicha 2014年04月30日 11:17 2621

R语言常用函数之aggregate

u011219650 2014年11月22日 11:53 11662

aggregate函数应该是数据处理中常用到的函数，简单说有点类似sql语言中的group by，可以按照要求把数据打组聚合，然后对聚合以后的数据进行加和、求平均等各种操作。 x=da...

如何学习大数据



联系我们

请扫描二维码联系

webmaster@

400-660-0108

QQ客服 客