详细探究Spark的shuffle实现

architecture (10) (/categories.html#architecture-ref)

spark 11 (/tags.html#spark-ref) shuffle 1 (/tags.html#shuffle-ref)

mapreduce 2 (/tags.html#mapreduce-ref) cloud 10 (/tags.html#cloud-ref)

04 January 2014

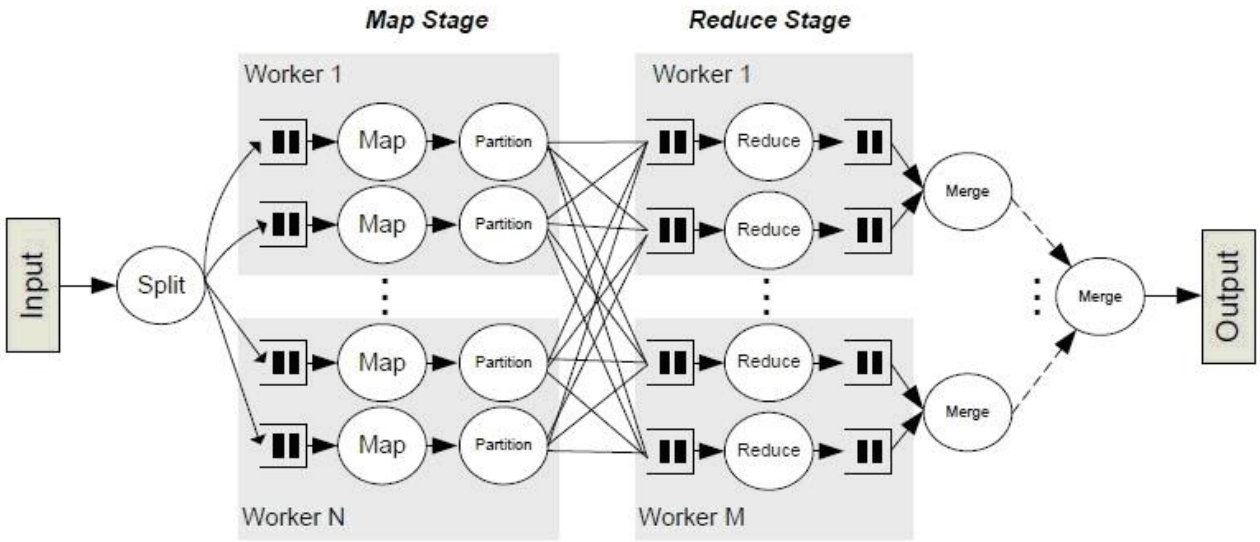
Background

在MapReduce框架中，shuffle是连接Map和Reduce之间的桥梁，Map的输出要用到Reduce中必须经过shuffle这个环节，shuffle的性能高低直接影响了整个程序的性能和吞吐量。Spark作为MapReduce框架的一种实现，自然也实现了shuffle的逻辑，本文就深入研究Spark的shuffle是如何实现的，有什么优缺点，与Hadoop MapReduce的shuffle有什么不同。

Shuffle

Shuffle是MapReduce框架中的一个特定的phase，介于Map phase和Reduce phase之间，当Map的输出结果要被Reduce使用时，输出结果需要按key哈希，并且分发到每一个Reducer上去，这个过程就是shuffle。由于shuffle涉及到了磁盘的读写和网络的传输，因此shuffle性能的高低直接影响到了整个程序的运行效率。

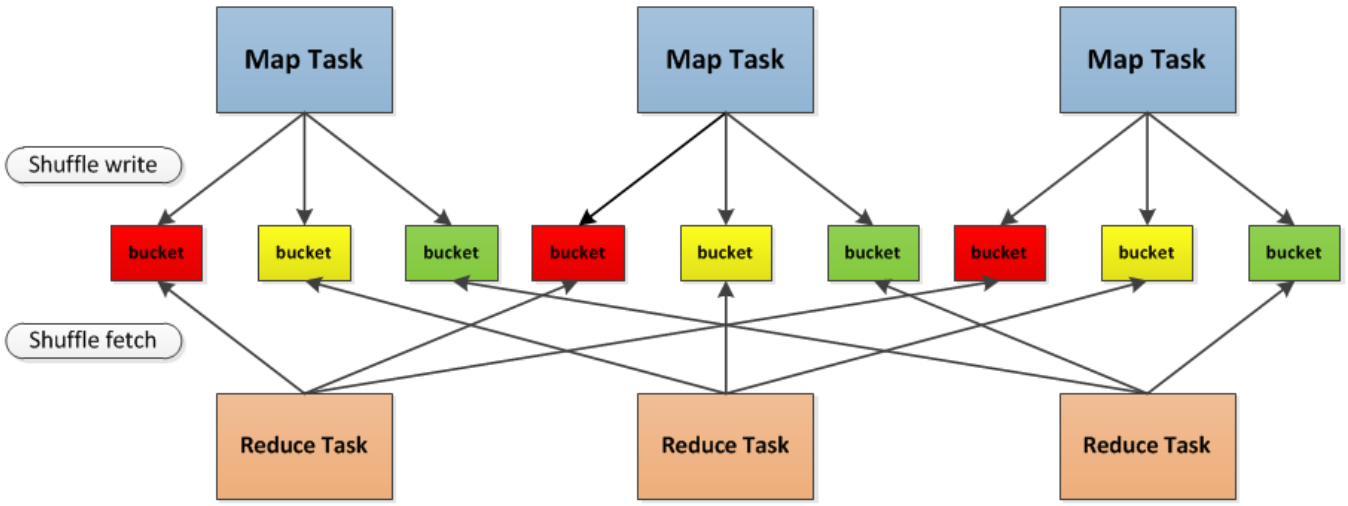
下面这幅图清晰地描述了MapReduce算法的整个流程，其中shuffle phase是介于Map phase和Reduce phase之间。



概念上shuffle就是一个沟通数据连接的桥梁，那么实际上shuffle这一部分是如何实现的的呢，下面我们就以Spark为例讲一下shuffle在Spark中的实现。

Spark Shuffle进化史

先以图为例简单描述一下Spark中shuffle的整一个流程：



- 首先每一个Mapper会根据Reducer的数量创建出相应的bucket，bucket的数量是$M \times R$，其中M是Map的个数，R是Reduce的个数。
- 其次Mapper产生的结果会根据设置的partition算法填充到每个bucket中去。这里的partition算法是可以自定义的，当然默认的算法是根据key哈希到不同的bucket中去。
- 当Reducer启动时，它会根据自己task的id和所依赖的Mapper的id从远端或是本地的block manager中取得相应的bucket作为Reducer的输入进行处理。

这里的bucket是一个抽象概念，在实现中每个bucket可以对应一个文件，可以对应文件的一部分或是其他等。

接下来我们分别从**shuffle write**和**shuffle fetch**这两块来讲述一下Spark的shuffle进化史。

Shuffle Write

在Spark 0.6和0.7的版本中，对于shuffle数据的存储是以文件的方式存储在block manager中，与`rdd.persist(StorageLevel.DISk_ONLY)`采取相同的策略，可以参看：

```
override def run(attemptId: Long): MapStatus = {
  val numOutputSplits = dep.partitioner.numPartitions

  ...
    // Partition the map output.
    val buckets = Array.fill(numOutputSplits)(new ArrayBuffer[(Any, Any)])
    for (elem <- rdd.iterator(split, taskContext)) {
      val pair = elem.asInstanceOf[(Any, Any)]
      val bucketId = dep.partitioner.getPartition(pair._1)
      buckets(bucketId) += pair
    }

    ...

    val blockManager = SparkEnv.get.blockManager
    for (i <- 0 until numOutputSplits) {
      val blockId = "shuffle_" + dep.shuffleId + "_" + partition + "_" + i
      // Get a Scala iterator from Java map
      val iter: Iterator[(Any, Any)] = buckets(i).iterator
      val size = blockManager.put(blockId, iter, StorageLevel.DISK_ONLY, false)
      totalBytes += size
    }
  ...
}
```

我已经将一些干扰代码删去。可以看到Spark在每一个Mapper中为每个Reducer创建一个bucket，并将RDD计算结果放进bucket中。需要注意的是每个bucket是一个`ArrayBuffer`，也就是说Map的输出结果是会先存储在内存。

接着Spark会将ArrayBuffer中的Map输出结果写入block manager所管理的磁盘中，这里文件的命名方式为：`shuffle_ + shuffle_id + "_" + map partition id + "_" + shuffle partition id`。

早期的shuffle write有两个比较大的问题：

1. Map的输出必须先全部存储到内存中，然后写入磁盘。这对内存是一个非常大的开销，当内存不足以存储所有的Map output时就会出现OOM。
2. 每一个Mapper都会产生Reducer number个shuffle文件，如果Mapper个数是1k，Reducer个数也是1k，那么就会产生1M个shuffle文件，这对于文件系统是一个非常大的负担。同时在shuffle数据量不大而shuffle文件又非常多的情况下，随机写也会严重降低IO的性能。

在Spark 0.8版本中，shuffle write采用了与RDD block write不同的方式，同时也为shuffle write单独创建了`ShuffleBlockManager`，部分解决了0.6和0.7版本中遇到的问题。

首先我们来看一下Spark 0.8的具体实现：

```
override def run(attemptId: Long): MapStatus = {

  ...

  val blockManager = SparkEnv.get.blockManager
  var shuffle: ShuffleBlocks = null
  var buckets: ShuffleWriterGroup = null

  try {
    // Obtain all the block writers for shuffle blocks.
    val ser = SparkEnv.get.serializerManager.get(dep.serializerClass)
    shuffle = blockManager.shuffleBlockManager.forShuffle(dep.shuffleId, numOutputSplits, ser)
    buckets = shuffle.acquireWriters(partition)

    // Write the map output to its associated buckets.
    for (elem <- rdd.iterator(split, taskContext)) {
      val pair = elem.asInstanceOf[Product2[Any, Any]]
      val bucketId = dep.partitioner.getPartition(pair._1)
      buckets.writers(bucketId).write(pair)
    }

    // Commit the writes. Get the size of each bucket block (total block size).
    var totalBytes = 0L
    val compressedSizes: Array[Byte] = buckets.writers.map { writer:   BlockObjectWriter =>
      writer.commit()
      writer.close()
      val size = writer.size()
      totalBytes += size
      MapOutputTracker.compressSize(size)
    }

    ...

  } catch { case e: Exception =>
    // If there is an exception from running the task, revert the partial writes
    // and throw the exception upstream to Spark.
    if (buckets != null) {
      buckets.writers.foreach(_.revertPartialWrites())
    }
    throw e
  } finally {
    // Release the writers back to the shuffle block manager.
    if (shuffle != null && buckets != null) {
      shuffle.releaseWriters(buckets)
    }
    // Execute the callbacks on task completion.
    taskContext.executeOnCompleteCallbacks()
    }
  }
}
```

在这个版本中为shuffle write添加了一个新的类`ShuffleBlockManager`，由`ShuffleBlockManager`来分配和管理bucket。同时`ShuffleBlockManager`为每一个bucket分配一个`DiskObjectWriter`，每个write handler拥有默认100KB的缓存，使用这个write handler将Map output写入文件中。可以看到现在的写入方式变为`buckets.writers(bucketId).write(pair)`，也就是说Map output的key-value pair是逐个写入到磁盘而不是预先把所有数据存储在内存中在整体flush到磁盘中去。

`ShuffleBlockManager`的代码如下所示：

```
private[spark]
class ShuffleBlockManager(blockManager: BlockManager) {

  def forShuffle(shuffleId: Int, numBuckets: Int, serializer: Serializer): ShuffleBlocks = {
    new ShuffleBlocks {
      // Get a group of writers for a map task.
      override def acquireWriters(mapId: Int): ShuffleWriterGroup = {
        val bufferSize = System.getProperty("spark.shuffle.file.buffer.kb", "100").toInt * 1024
        val writers = Array.tabulate[BlockObjectWriter](numBuckets) { bucketId =>
          val blockId = ShuffleBlockManager.blockId(shuffleId, bucketId, mapId)
          blockManager.getDiskBlockWriter(blockId, serializer, bufferSize)
        }
        new ShuffleWriterGroup(mapId, writers)
      }

      override def releaseWriters(group: ShuffleWriterGroup) = {
        // Nothing really to release here.
      }
    }
  }
}
```

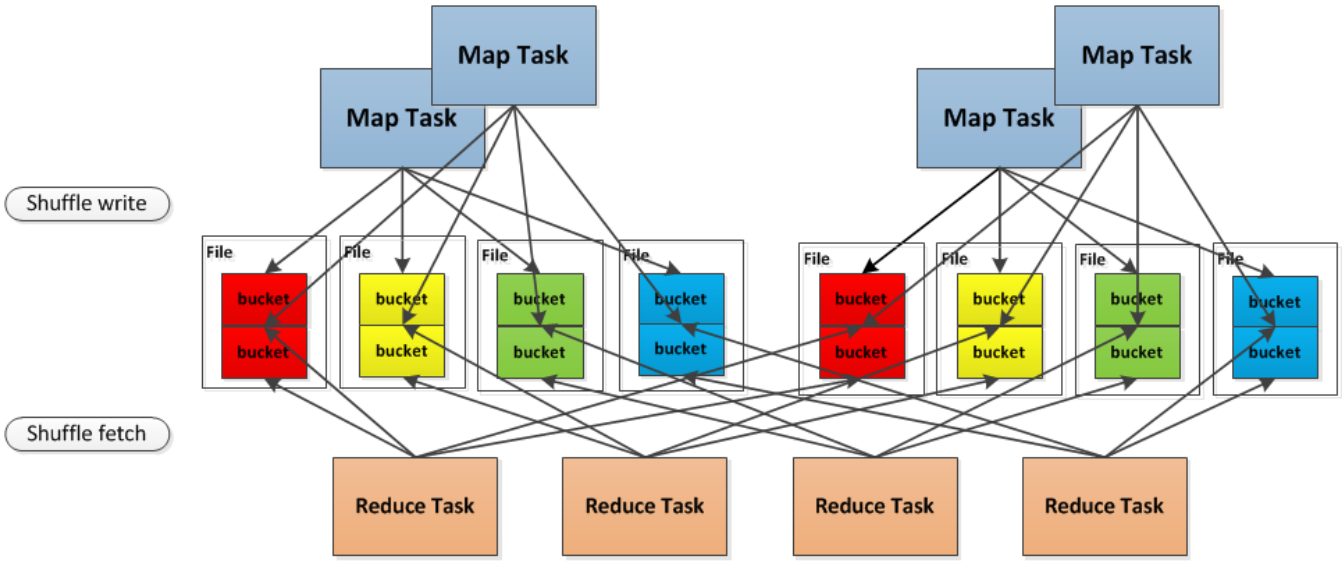
Spark 0.8显著减少了shuffle的内存压力，现在Map output不需要先全部存储在内存中，再flush到硬盘，而是record-by-record写入到磁盘中。同时对于shuffle文件的管理也独立出新的`ShuffleBlockManager`进行管理，而不是与rdd cache文件在一起了。

但是这一版Spark 0.8的shuffle write仍然有两个大的问题没有解决：

- 首先依旧是shuffle文件过多的问题，shuffle文件过多一是会造成文件系统的压力过大，二是会降低IO的吞吐量。
- 其次虽然Map output数据不再需要预先在内存中evaluate显著减少了内存压力，但是新引入的`DiskObjectWriter`所带来的buffer开销也是一个不容小视的内存开销。假定我们有1k个Mapper和1k个Reducer，那么就会有1M个bucket，于此同时就会有1M个write handler，而每一个write handler默认需要100KB内存，那么总共需要100GB的内存。这样的话仅仅是buffer就需要这么多的内存，内存的开销是惊人的。当然实际情况下这1k个Mapper是分时运行的话，所需的内存就只有`cores * reducer numbers * 100KB`大小了。但是reducer数量很多的话，这个buffer的内存开销也是蛮厉害的。

为了解决shuffle文件过多的情况，Spark 0.8.1引入了新的shuffle consolidation，以期显著减少shuffle文件的数量。

首先我们以图例来介绍一下shuffle consolidation的原理。



假定该job有4个Mapper和4个Reducer，有2个core，也就是能并行运行两个task。我们可以算出Spark的shuffle write共需要16个bucket，也就有了16个write handler。在之前的Spark版本中，每一个bucket对应的是一个文件，因此在这里会产生16个shuffle文件。

而在shuffle consolidation中每一个bucket并非对应一个文件，而是对应文件中的一个segment，同时shuffle consolidation所产生的shuffle文件数量与Spark core的个数也有关系。在上面的图例中，job的4个Mapper分为两批运行，在第一批2个Mapper运行时会申请8个bucket，产生8个shuffle文件；而在第二批Mapper运行时，申请的8个bucket并不会再产生8个新的文件，而是追加写到之前的8个文件后面，这样一共就只有8个shuffle文件，而在文件内部这有16个不同的segment。因此从理论上讲shuffle consolidation所产生的shuffle文件数量为$C \times R$，其中C是Spark集群的core number，R是Reducer的个数。

需要注意的是当 $M = C$时shuffle consolidation所产生的文件数和之前的实现是一样的。

Shuffle consolidation显著减少了shuffle文件的数量，解决了之前版本一个比较严重的问题，但是writer handler的buffer开销过大依然没有减少，若要减少writer handler的buffer开销，我们只能减少Reducer的数量，但是这又会引入新的问题，下文将会有详细介绍。

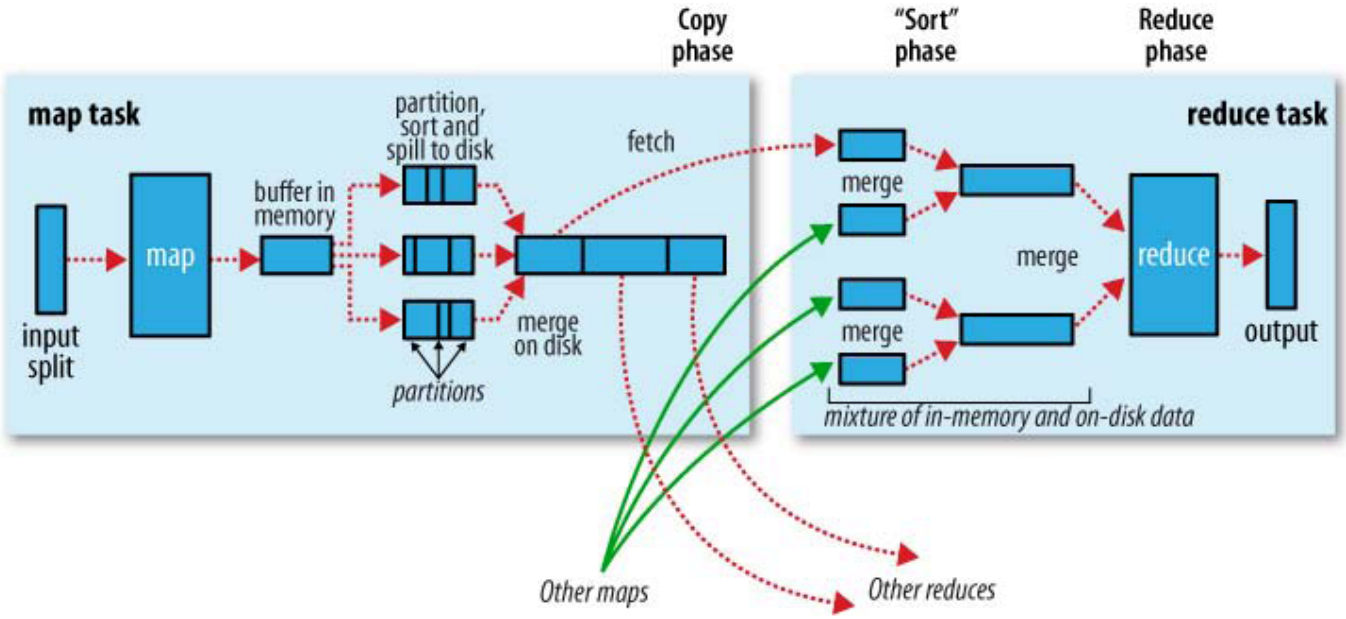
讲完了shuffle write的进化史，接下来要讲一下shuffle fetch了，同时还要讲一下Spark的aggregator，这一块对于Spark实际应用的性能至关重要。

Shuffle Fetch and Aggregator

Shuffle write写出去的数据要被Reducer使用，就需要shuffle fetcher将所需的数据fetch过来，这里的fetch包括本地和远端，因为shuffle数据有可能一部分是存储在本地的。Spark对shuffle fetcher实现了两套不同的框架：NIO通过socket连接去fetch数据；OIO通过netty server去fetch数据。分别对应的类是`BasicBlockFetcherIterator`和`NettyBlockFetcherIterator`。

在Spark 0.7和更早的版本中，只支持`BasicBlockFetcherIterator`，而`BasicBlockFetcherIterator`在shuffle数据量比较大的情况下performance始终不是很好，无法充分利用网络带宽，为了解决这个问题，添加了新的shuffle fetcher来试图取得更好的性能。对于早期shuffle性能的评测可以参看Spark usergroup (https://groups.google.com/forum/#!msg/shark-users/IHOb2u5HXSk/huTWyosI1n4J)。当然现在`BasicBlockFetcherIterator`的性能也已经好了很多，使用的时候可以对这两种实现都进行测试比较。

接下来说一下aggregator。我们都知道在Hadoop MapReduce的shuffle过程中，shuffle fetch过来的数据会进行merge sort，使得相同key下的不同value按序归并到一起供Reducer使用，这个过程可以参看下图：



所有的merge sort都是在磁盘上进行的，有效地控制了内存的使用，但是代价是更多的磁盘IO。

那么Spark是否也有merge sort呢，还是以别的方式实现，下面我们就细细说明。

首先虽然Spark属于MapReduce体系，但是对传统的MapReduce算法进行了一定的改变。Spark假定在大多数用户的case中，shuffle数据的sort不是必须的，比如word count，强制地进行排序只会使性能变差，因此Spark并不在Reducer端做merge sort。既然没有merge sort那Spark是如何进行reduce的呢？这就要说到aggregator了。

aggregator本质上是一个hashmap，它是以map output的key为key，以任意所要combine的类型为value的hashmap。当我们在做word count reduce计算count值的时候，它会将shuffle fetch到的每一个key-value pair更新或是插入到hashmap中(若在hashmap中没有查找到，则插入其中；若查找到则更新value值)。这样就不需要预先把所有的key-value进行merge sort，而是来一个处理一个，省下了外部排序这一步骤。但同时需要注意的是reducer的内存必须足以存放这个partition的所有key和count值，因此对内存有一定的要求。

在上面word count的例子中，因为value会不断地更新，而不需要将其全部记录在内存中，因此内存的使用还是比较少的。考虑一下如果是group by key这样的操作，Reducer需要得到key对应的所有value。在Hadoop MapReduce中，由于有了merge sort，因此给予Reducer的数据已经是group by key了，而Spark没有这一步，因此需要将key和对应的value全部存放在hashmap中，并将value合并成一个array。可以想象为了能够存放所有数据，用户必须确保每一个partition足够小到内存能够容纳，这对于内存是一个非常严峻的考验。因此Spark文档中建议用户涉及到这类操作的时候尽量增加partition，也就是增加Mapper和Reducer的数量。

增加Mapper和Reducer的数量固然可以减小partition的大小，使得内存可以容纳这个partition。但是我们在shuffle write中提到，bucket和对应于bucket的write handler是由Mapper和Reducer的数量决定的，task越多，bucket就会增加的更多，由此带来write handler所需的buffer也会更多。在一方面我们为了减少内存的使用采取了增加task数量的策略，另一方面task数量增多又会带来buffer开销更大的问题，因此陷入了内存使用的两难境地。

为了减少内存的使用，只能将aggregator的操作从内存移到磁盘上进行，Spark社区也意识到了Spark在处理数据规模远远大于内存大小时所带来的问题。因此PR303 (https://github.com/apache/incubator-spark/pull/303)提供了外部排序的实现方案，相信在Spark 0.9 release的时候，这个patch应该能merge进去，到时候内存的使用量可以显著地减少。

End

本文详细地介绍了Spark的shuffle实现是如何进化的，以及遇到问题解决问题的过程。shuffle作为Spark程序中很重要的一个环节，直接影响了Spark程序的性能，现如今的Spark版本虽然shuffle实现还存在着种种问题，但是相比于早期版本，已经有了很大的进步。开源代码就是如此不停地迭代推进，随着Spark的普及程度

越来越高，贡献的人越来越多，相信后续的版本会有更大的提升。

← Previous (/algorithm/2014/01/02/simrank-mapreduce-survey) | Archive (/archive.html)
Next → (/architecture/2015/08/22/spark-dynamic-allocation-investigation)

haidao • 2 years ago

感谢作者的分析。对于spark 的shuffle，我有两个疑问
1 你说的‘也就是说Map output的key-value pair是逐个写入到磁盘而不是预先把所有数据存储在内存中在整体flush到磁盘中去。’这样一个一个写磁盘会不会性能太低，我记得hadoop里会设置一个10M的buffer，然后依次写入磁盘，并且这个是可以配置的，当处理不同数据时配置不同的buffer能够极大提高系统性能。

2 分析中貌似没有对map中bucket的分割函数有具体介绍 难道是简单的hash（key/reduce_num）？ 在hadoop里hash函数为了使数据分割更均匀，会有不同策略的hash函数，记得有个比较高级的是 根据数据的采样结果来设计hash函数的具体实现。

总之感觉hadoop在一些细节优化上做的比较好。我感觉spark 很厉害的地方在于
1.内存优先的原则，比如你在上一篇文章中所说“容错所引入的昂贵数据实体化”，各种mapredcue的中间结果写磁盘确实很坑爹。
2 DAG 确实很优秀，可以避免很多不必要的task
3 scala 写代码确实比较爽，减少了很多不必要的代码。比java好了不知多少倍

另外spark的设计思路很清晰，看了作者介绍了的几篇文章，很容易就懂了一个大概。感谢作者的分享

1 • Reply • Share ›

jerryshao Mod ➚ haidao • 2 years ago

1. Spark对于Map output的每一个bucket是有默认100KB的缓存的，因此也不是一个个直接flush到磁盘中去的。由于Spark的shuffle实现与Hadoop不同，所以就没有类似于Hadoop的ring buffer来缓存Map output.
2. Spark当然也有不同的partition策略，并且也可以实现自己的partition策略，当然默认是根据key的Hash来分割。

2 • Reply • Share ›

wfeng1982@163.com ➚ jerryshao • 2 years ago

ringbuffer在mr中 主要用来解除写内存线程和spill到磁盘的线程的并发操作，

和spark shuffle方式与hadoop不同并没有什么关系。我现在想确定的是100KB满

了时是wait 到写磁盘完成么？

Reply • Share ›

jerryshao Mod → wfeng1982@163.com • 2 years ago

100KB是BufferedOutputStream所维护的，Spark只负责把数据写入流中，至于流满了是否阻塞应该是java的策略。不过我觉得如果数据不能立即写入磁盘而造成buffer满应该会造成写阻塞。

Reply • Share ›

王猛 • 2 years ago

内容很足，谢谢分享，果然至少在shuffle这一块并不是spark快的主因素，谢谢你的好文！

Reply • Share ›

松 张 • 2 years ago

看了几篇您的文章，收益匪浅。想问一下，在shuffle阶段，spark比mapreduce快就在于少了一个外部排序对么，shuffle的前半段，就是map将其写进磁盘两个是一样的对吗？，还有在数据量不大的情况下，spark还会将map的输出写进磁盘么，谢谢

Reply • Share ›

jerryshao Mod → 松 张 • 2 years ago

Hi 张松，相比与MR，Spark在shuffle阶段是不需要排序的，在某些case中这种优势比较明显；但是对于像streaming join这样的case，必须要将shuffle数据排序，这时候优势就不明显了。其次相比与MR的shuffle实现（包括prefetch，file combine），Spark的shuffle实现还比较原始，个人觉得Spark的快更大程度上是由于DAG优化和中间数据cache而带来的，shuffle尤其是大数据的shuffle Spark不一定有MR快。

其次两者都需要将map的输出写进磁盘，但是实现方式不太一样。还有不管什么情况下Spark都需要将map的输出写进磁盘，无论数据的大小。

Reply • Share ›

Zheng Han → jerryshao • 2 years ago

博主，您好！
最近在研究spark，跑graphx中的triangle counting恰好就遇到了这个mapshuffle输出结果过多导致OOM的问题，有些具体问题想和您学习交流一下，可以留个联系方式吗？谢谢！

4 Reply • Share ›

松 张 → jerryshao • 2 years ago

哦，非常感谢。我之所以问这个是看到那个在ml计算中第一代也比mapreduce快，所以才问的上述问题。谢谢了

1 Reply • Share ›

qianlan → jerryshao • 2 years ago

博主，你说的spark的快比如：DAG优化和中间数据cache。中间数据cache指的是：Map output的每一个bucket有默认100KB缓存么？还是指什么呢？

Reply • Share ›

mingjie • a year ago

写的非常棒 结合1.3代码走了一遍文档，有些地方已经作了修改，比如说shuffler writer and reader 有了两种实现，一种是hash based，一种是sort based, 但是基本的架构没变化的。回复上面的问题，spark 快的一个大的原因是定义了partition 的依赖关系，这样就减少了没有必要的shuffle stage, 这样就减少了HDFS 文件的读写。根据最新的一个论文 EDBT15，hadoop mapreduce 的一个大问题就是 hdfs i/o 最慢。 具体原因不详，自己感觉可能是HDFS 文件的写文件三份备份的策略，以及从远程读取文件的原因。总之，jerry shao and jerry lead 的介绍都写得很好，非常感谢了。

  • Reply • Share ›

ALSO ON **JERRY SHAO'S HOMEPAGE**

Spark Streaming Introduction

1 comment • 3 years ago

Avat **ruson** — 大赞

Spark源码分析之-deploy模块

3 comments • 3 years ago

Avat **Huangdong Meng** — 明白～ 期待大神的更多大作哈～ 比如shark～ RDD的部分复杂的operator的实现 等data processing层的讲解

Investigation of Dynamic Allocation in Spark

1 comment • 8 months ago

Avat **Robert Towne** — Jerry, good information. I'd like to push to graphite (or any monitoring/graphing system) how the …

传统的MapReduce框架慢在那里

1 comment • 3 years ago

Avat **Gavin Zhang** — 学习了，写得非常好！

CATEGORIES

test (1) (/categories.html#test-ref)

architecture (10) (/categories.html#architecture-ref)

functional programming (1) (/categories.html#functional programming-ref)

algorithm (1) (/categories.html#algorithm-ref)

LINKS

阮一峰的网络日志 (http://www.ruanyifeng.com/blog/)

刘未鹏 (http://mindhacks.cn/)

酷壳 (http://coolshell.cn/)

BeiYuu.com (http://beiyuu.com/)

MY FAVORITES

(http://movie.douban.com/subject/1652587/)


(http://movie.douban.com/subject/2043546/)

(https://music.douban.com/subject/1394568/)


(http://movie.douban.com/subject/1293764/)


(http://movie.douban.com/subject/3908423/)


(http://movie.douban.com/subject/1770547/)


(https://music.douban.com/subject/1427956/)


(http://movie.douban.com/subject/1810517/)

豆瓣douban.com (http://www.douban.com/)