Unlike Spark standalone and Mesos modes, in which the master’s address is specified in the --master parameter,  
in YARN mode the ResourceManager’s address is picked up from the Hadoop configuration. Thus,  
the --master parameter is yarn.

提交任务命令:

$ ./bin/spark-submit --class org.apache.spark.examples.SparkPi \

--master yarn \

--deploy-mode cluster \

--driver-memory 4g \

--executor-memory 2g \

--executor-cores 1 \

--queue thequeue \

lib/spark-examples\*.jar \

--conf spark.default.parallelism=300

--jars a.jar,b.jar

10

nohup spark-submit --master spark://hxf:7077

--executor-memory 30G

--executor-cores 24

--conf spark.default.parallelism=300

--class com.analysis.main.GenAppInfo

/home/hadoop/jar/analysis.jar $dst\_date > /home/hadoop/logs/genAppInfo.log &

yarn logs -applicationId application\_1511334364226\_0440

yarn application -kill appid//appid

**参数:**

spark.default.parallelizm：并行度问题，如果不设置这个参数，Spark 会跟据 HDFS 中 Block 的个数去设置这一个数量，原理是默应每个 Block 会对应一个 Task，默应情况下，如果数据量不是太多就不可以充份利用 executor 设置的资源，就会浪费了资源。建义设置为 100个，最好 700个左右。Spark官方的建义是每一个 Core 负责 2-3 个 Task。

spark.storage.memoryFraction：默应占用 60％，如果计算比较依赖于历史数据则可以调高该参数，当如果计算比较依赖 Shuffle 的话则需要降低该比例。

spark.shuffle.memoryFraction：默应占用 20％，如果计算比较依赖 Shuffle 的话则需要调高该比例。

spark.shuffle.safetyFraction=0.8

**Spark.shuffle.manager**=hash,sort,tungsten-sort

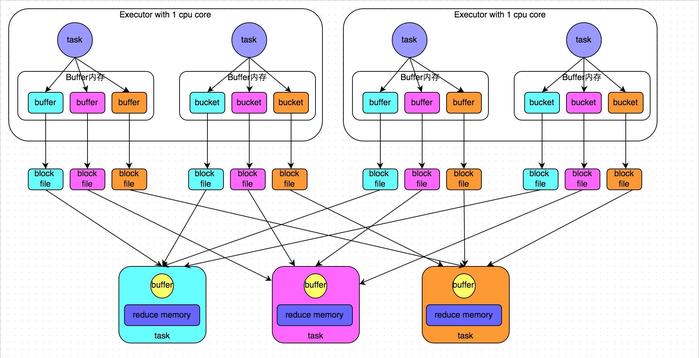
**HashShuffleManager**：shuffle write阶段，默认Mapper阶段会为Reducer阶段的每一个Task单独创建一个文件来保存该Task中要使用的数据。

优点：就是操作数据简单。

缺点：大量中间文件，大量io，容易OOM

优化: spark.shuffle.consolidateFiles=true，优化之后每个exucutor产生reduce task数量的文件

CPU core的数量 \* 下一个stage的task数量



**SortShuffleManager**

在Mapper中的每一个ShuffleMapTask中产生两个文件：Data文件和Index文件，产生任务数 2 \* shufflemaptask

Sorted-Based Shuffle 在 **Mapper 端是排序的**，包括 partition 的排序和每个 partition 内部元素的排序！但在 **Reducer 端是没有进行排序**，所以 Job 的结果默认不是排序的。Sorted-Based Shuffle 采用了 **Tim-Sort 排序算法**，好处是可以极为高效的使用 Mapper 端的排序成果全局排序Sorted-Based Shuffle 的核心是借助于 ExternalSorter 把每个 ShuffleMapTask 的输出排序到一个文件中(FileSegmentGroup)

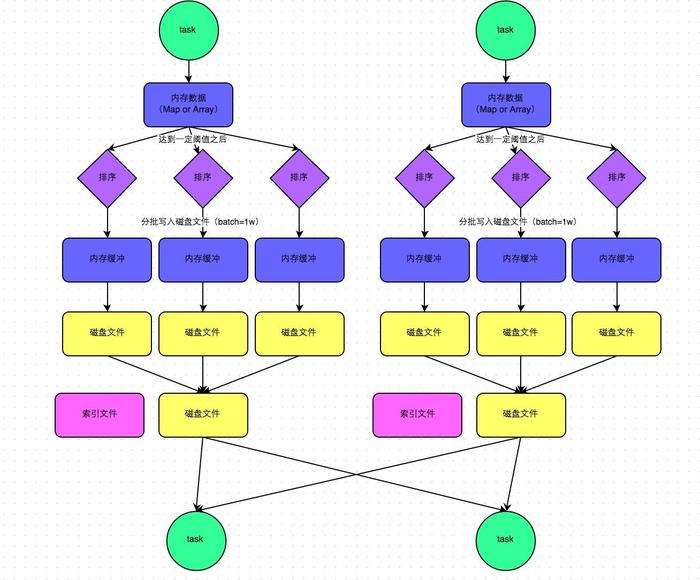
Bypass机制:触发条件1：shuffle map task数量小于spark.shuffle.sort.bypassMergeThreshold参数的值,默认200,此时会使用类似hashshuffle将分区文件写为单独文件，最后进行合并。这样通过去除Sort步骤来加快处理速度，代价是需要并发打开多个文件，所以内存消耗量增加。触发条件2：不是聚合类的shuffle算子（比如reduceByKey）。

优点:中间文件数量减少，节省io,内存

缺点:shufflemaptask如果数量很多，依旧会产生很多小文件

若需要分片内排序，需进行mapper端和reducer端两次排序，默认reduce端不排序。

不需要排序的操作也会执行map端排序，浪费计算资源



**Tungston-sort**：机制与sort类似，但是用到了对外内存，增加了对内存和cpu的优化

**Catalyst：是一种解析引擎，不仅是sql解析引擎。采用数据结构，使用解析规则解析为语法树。**

**Shuffle相关参数:**

**[spark.shuffle.file.buffer](https://www.jianshu.com/p/98a1d67bc226" \t "https://www.jianshu.com/p/_blank)**默认值：32k

参数说明：该参数用于设置shuffle write task的BufferedOutputStream的buffer缓冲大小。将数据写到磁盘文件之前，会先写入buffer缓冲中，待缓冲写满之后，才会溢写到磁盘。

调优建议：如果作业可用的内存资源较为充足的话，可以适当增加这个参数的大小（比如64k），从而减少shuffle write过程中溢写磁盘文件的次数，也就可以减少磁盘IO次数，进而提升性能。在实践中发现，合理调节该参数，性能会有1%~5%的提升。

**[spark.reducer.maxSizeInFlight](" \t "_blank)** 默认值：48m

参数说明：该参数用于设置shuffle read task的buffer缓冲大小，而这个buffer缓冲决定了每次能够拉取多少数据。

调优建议：如果作业可用的内存资源较为充足的话，可以适当增加这个参数的大小（比如96m），从而减少拉取数据的次数，也就可以减少网络传输的次数，进而提升性能。在实践中发现，合理调节该参数，性能会有1%~5%的提升。

**[spark.shuffle.io.maxRetries](" \t "_blank)** 默认值：3

参数说明：shuffle read task从shuffle write task所在节点拉取属于自己的数据时，如果因为网络异常导致拉取失败，是会自动进行重试的。该参数就代表了可以重试的最大次数。如果在指定次数之内拉取还是没有成功，就可能会导致作业执行失败。

调优建议：对于那些包含了特别耗时的shuffle操作的作业，建议增加重试最大次数（比如60次），以避免由于JVM的full gc或者网络不稳定等因素导致的数据拉取失败。在实践中发现，对于针对超大数据量（数十亿~上百亿）的shuffle过程，调节该参数可以大幅度提升稳定性。

**[spark.shuffle.io.retryWait](https://www.jianshu.com/p/98a1d67bc226" \t "https://www.jianshu.com/p/_blank)** 默认值：5s

参数说明：具体解释同上，该参数代表了每次重试拉取数据的等待间隔，默认是5s。

调优建议：建议加大间隔时长（比如60s），以增加shuffle操作的稳定性。

**[spark.shuffle.consolidateFiles](" \t "_blank)** 默认值：false

参数说明：如果使用HashShuffleManager，该参数有效。如果设置为true，那么就会开启consolidate机制，会大幅度合并shuffle write的输出文件，对于shuffle read task数量特别多的情况下，这种方法可以极大地减少磁盘IO开销，提升性能。

调优建议：如果的确不需要SortShuffleManager的排序机制，那么除了使用bypass机制，还可以尝试将spark.shffle.manager参数手动指定为hash，使用HashShuffleManager，同时开启consolidate机制。在实践中尝试过，发现其性能比开启了bypass机制的SortShuffleManager要高出10%~30%

**spark.ui.killEnabled:true**

运行在web UI中杀死stage和相应的job

**spark.ui.port:4040**

你的应用程序dashboard的端口。显示内存和工作量数据

**spark.rdd.compress:true**

是否压缩序列化的RDD分区。在花费一些额外的CPU时间的同时节省大量的空

**spark.broadcast.compress:true**

在发送广播变量之前是否压缩

**spark.io.compression.codec:snappy**

压缩诸如RDD分区、广播变量、shuffle输出等内部数据的编码解码器。默认情况下，Spark提供了三种选择：lz4、lzf和snappy，你也可以用完整的类名来制定

**spark.kryo.classesToRegister:(none)**

如果你用Kryo序列化，给定的用逗号分隔的自定义类名列表表示要注册的类

**spark.kryoserializer.buffer.max.mb:64**

Kryo序列化缓存允许的最大值。这个值必须大于你尝试序列化的对象

**spark.serializer:org.apache.spark.serializer.JavaSerializer**

序列化对象使用的类。默认的Java序列化类可以序列化任何可序列化的java对象但是它很慢。所有我们建议用org.apache.spark.serializer.KryoSerializer

**spark.cleaner.ttl:(infinite)**

spark记录任何元数据（stages生成、task生成等）的持续时间。定期清理可以确保将超期的元数据丢弃，这在运行长时间任务是很有用的，如运行7\*24的sparkstreaming任务。RDD持久化在内存中的超期数据也会被清理

**spark.default.parallelism**

本地模式：机器核数；Mesos：8；其他：max(executor的core，2)

如果用户不设置，系统使用集群中运行shuffle操作的默认任务数（groupByKey、 reduceByKey等）

**spark.files.fetchTimeout:60**

driver 程序获取通过SparkContext.addFile()添加的文件时的超时时间，单位秒

**spark.storage.unrollFraction:0.2**

Fraction of spark.storage.memoryFraction to use for unrolling blocks in memory.

**spark.tachyonStore.baseDir:System.getProperty(“java.io.tmpdir”)**

Tachyon File System临时目录

**spark.tachyonStore.url:tachyon://localhost:19998**

Tachyon File System URL

**spark.speculation.multiplier:1.5**

比已完成Task的运行速度中位数慢多少倍才启用推测

**spark.locality.wait:3000**

以下几个参数是关于Spark数据本地性的。本参数是以毫秒为单位启动本地数据task的等待时间，如果超出就启动下一本地优先级别 的task。该设置同样可以应用到各优先级别的本地性之间（本地进程 -> 本地节点 -> 本地机架 -> 任意节点 ），当然，也可以通过spark.locality.wait.node等参数设置不同优先级别的本地性

**当你运行Spark Standalone Mode或者Spark on Mesos模式时，你可以通过Spark History Server来查看job运行情况。**

**Kryo序列化**

1. val sparkConf = **new** SparkConf()
2. /\*.setMaster("local")\*/
3. .setAppName("Test")
4. .set("spark.serializer", "org.apache.spark.serializer.KryoSerializer")
5. /\*.set("spark.kryo.registrationRequired", "true")\*/
6. .registerKryoClasses(Array(
7. classOf[Array[String]],
8. classOf[util.HashMap[String, String]],
9. classOf[util.ArrayList[String]],
10. classOf[MyTest]
11. ))

**自定义累加器：累加器要在action算子后才能取到值，因为是lazy计算模式。下列计算结果 accum1:9 accum2:18，因为两个action算子，map任务执行了两遍，对应的累加器也累加了两次，可通过缓存来避免。**

1. val accum = sc.longAccumulator("longAccum")
2. val numberRDD = sc.parallelize(Array(1,2,3,4,5,6,7,8,9),2).map(n=>{
3. accum.add(1L)
4. n+1
5. })
6. numberRDD.count
7. println("accum1:"+accum.value)
8. numberRDD.reduce(\_+\_)
9. println("accum2: "+accum.value)
10. //修改之后
11. numberRDD.cache().count
12. println("accum1:"+accum.value)
13. numberRDD.reduce(\_+\_)
14. println("accum2: "+accum.value)

**2.X之前implements AccumulatorParam<String>**

**2.X之后extends AccumulatorV2[String, String]**

### **value方法：获取累加器中的值**

### **merge方法：该方法特别重要，一定要写对，这个方法是各个task的累加器进行合并的方法（下面介绍执行流程中将要用到）**

### **iszero方法：判断是否为初始值**

### **reset方法：重置累加器中的值**

### **copy方法：拷贝累加器**

### **spark中累加器的执行流程：**

### **首先有几个task，spark engine就调用copy方法拷贝几个累加器（不注册的），然后在各个task中进行累加（注意在此过程中，被最初注册的累加器的值是不变的），执行最后将调用merge方法和各个task的结果累计器进行合并（此时被注册的累加器是初始值）**

**Spark 2.1.0 新型 JVM Heap 分成三个部份：Reserved Memory、User Memory 和 Spark Memory。**

1. Reserved Memory：默认都是300MB
2. User Memory：写 Spark 程序中产生的临时数据或者是自己维护的一些数据结构也需要给予它一部份的存储空间，它占用的空间是 (Java Heap - Reserved Memory) x 25%(默认是25％，可以有参数供调优)
3. Spark Memeory：系统框架运行时需要使用的空间，这是从两部份构成的，分别是 Storage Memeory 和 Execution Memory。现在 Storage 和 Execution (Shuffle) 采用了 Unified 的方式共同使用了 (Heap Size - 300MB) x 75%，默认情况下 Storage 和 Execution 各占该空间的 50%。

Storage Memeory：相当于旧版本的 Storage 空间，在旧版本中 Storage 占了 54% 的 Heap 空间，这个空间会负责存储 Persist、Unroll 以及 Broadcast 的数据。

Execution Memeory：相当于旧版本的 Shuffle 空间，这个空间会负责存储 ShuffleMapTask 的数据。比如说从上一个 Stage 抓取数据和一些聚合的操作、等等。