容器中大数据业务性能优化技术报告

# 一、项目背景

Spark参数配置对引擎并行计算能力、系统稳定性具有重要影响。合理配置Spark参数，可以显著提高程序执行速度，提升引擎的稳定性。鉴于Spark参数数量的繁多、以及部分参数值域范围的广阔，无法通过枚举方式找出最优配置。本项目针对Spark参数配置问题进行研究，充分调研现有解决方案，分析现有方案各自优势，挖掘现有方案各自优势的根本原因，以最终提出的符合项目目标的Spark参数配置方案以及一系列过程性研究成果。调研结果表明，现有解决方案主要分为以下三类：

1）基于搜索的配置优化方法

2）基于机器学习的配置优化方法

3）基于规则的配置优化方法

具体调研结果见文档《容器中程序特征分析业界调研报告》。此外，在充分调研和深入分析的基础上，我们提出了一系列Spark性能调优方案，具体方案如第二章所述。

# SPARK性能调优

## 2.1、概述

根据论文调研结果，我们共设计了三套spark性能优化方案：搜索法、交互强度分析法、以及迭代建模法。在所有优化方案执行之前都需要对spark参数范围进行确定，因此我们研究出了spark配置参数范围确定方法。除此之外在研究TPC-DS时还设计了敏感query探测方法，以解决TPC-DS优化耗时过长的问题。基于我们设计的方案我们对数据量为100G的Wordcount（Wordcount-100G）、数据量为20G的Terasort（Terasort-20G）、以及数据量为21G的TPC-DS（tpcds-21G）进行了优化，优化结果显示我们设计的优化方案在上述三个用例上有较好的表现。对于Wordcount-100G、Terasort-20G和tpcds-21G，经过我们的优化得出的最优配置与基准配置对比，分别获得了12.4倍、9.2倍和1.9倍的优化效果。

## 2.2、调优场景介绍

### 2.2.1、实验环境

搭建以1个master节点、2个worker节点所构成的集群环境，所有节点统一使用Centos7 2009操作系统，各节点硬件配置如表2.1所示，集群组网图如图2.2所示，集群所使用软件版本如表2.3所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 节点名称 | CPU规格 | 内存 | 磁盘 |
| Master | Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz 20核 | 128G PC4 | 13T |
| Worker01 | Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz 20核 | 128G PC4 | 13T |
| Worker02 | Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz 20核 | 128G PC4 | 13T |

表2.1集群节点硬件配置

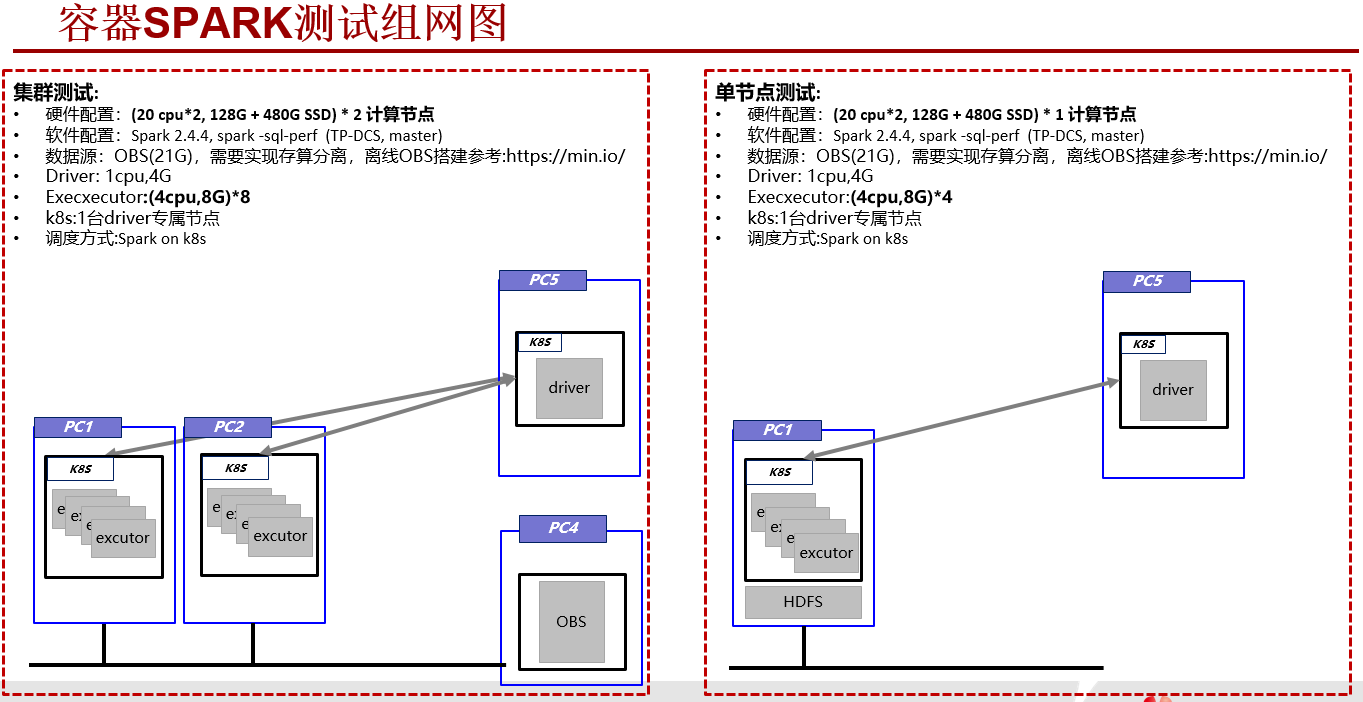


图2.2集群组网图

|  |  |
| --- | --- |
| 软件名称 | 版本号 |
| Spark | 2.4.4 |
| Docker | 20.10.5 |
| Kubernetes | 1.19.1 |
| Hibench | 6.0 |
| TPC-DS | 2.3.0 |

表2.3集群节点软件版本

### 2.2.2、测试用例

测试用例如表2.4所示，其中，Wordcount和Terasort来自Hibench。

|  |  |
| --- | --- |
| 测试用例 | 数据量大小 |
| Wordcount | 100G |
| Terasort | 20G |
| TPC-DS中全部103个query | 21G |

表2.4测试用例

## 2.3、优化方案设计

### 2.3.1、spark配置参数范围确定方法

项目所使用的参数范围确定方法可分为以下三个阶段：基于现有文献参考及参数含义分析的**参数范围预设置阶段**、基于**盒图分析**策略的**参数范围调整阶段**、以及基于**枚举策略**的**参数范围确定阶段**。参数范围确定方法处理流程如图2.5所示。

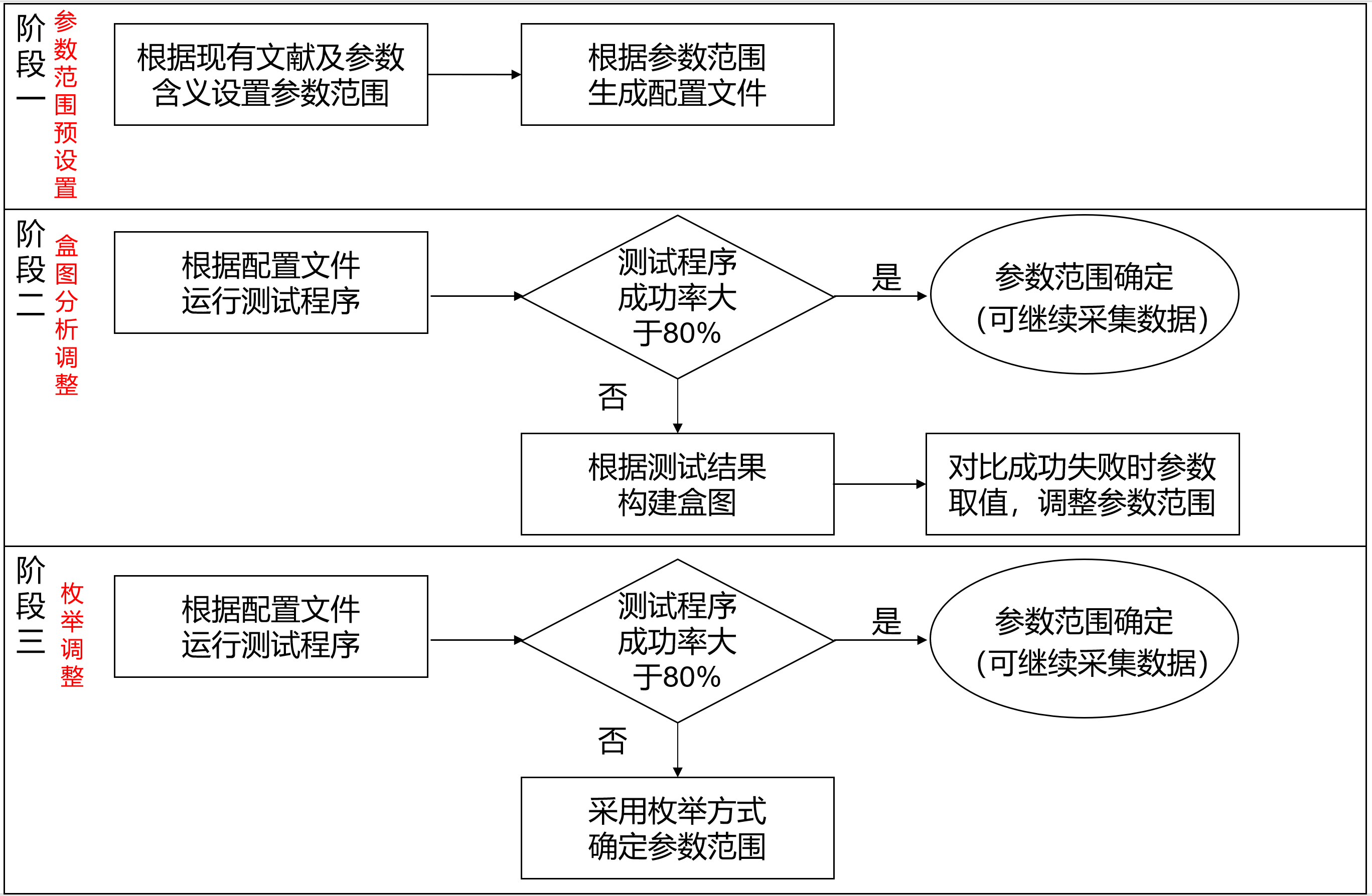


图2.5参数范围的确定

阶段一：对于现有文献已确定取值范围的参数，在文献参数范围的基础上进行轻微扩展（上下界扩展幅度为5%）；对于现有文献未确定取值范围的参数，根据参数含义和默认值设置参数范围。结合6篇文献中参数取值范围，确定了47个参数的范围，详见附件：论文中参数范围.xlsx。

阶段二：检验测试程序成功率时，程序运行时间不少于24小时；基于盒图的参数范围调整方案根据测试程序成功、失败时参数取值分布情况进行调整，具体而言，计算两种情况下的中位数差异，以计算结果与取值范围的商值作为基准，优先调整基准大的参数的取值范围，将参数范围向中位数差值方向调整两者中位数之差的30%，不对基准小于10%的参数范围进行调整，具体处理流程如图2.6所示。



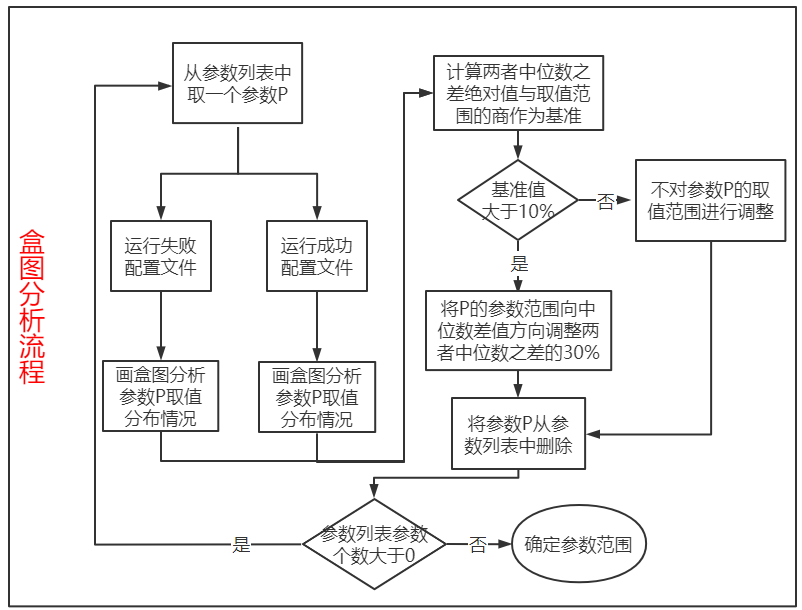


图2.6盒图分析流程

阶段三：检验测试程序成功率时，程序运行时间不少于12小时；枚举方式确定参数范围的主要思想为，根据报错日志选取相关参数，结合资源相关参数，构成重要参数集合，采用枚举方式逐一调整每个重要参数，直至获得测试程序可运行配置文件，最终以此配置为基准配置，根据参数含义，采用二分查找方式确定各参数取值范围，具体处理流程如图2.7所示。

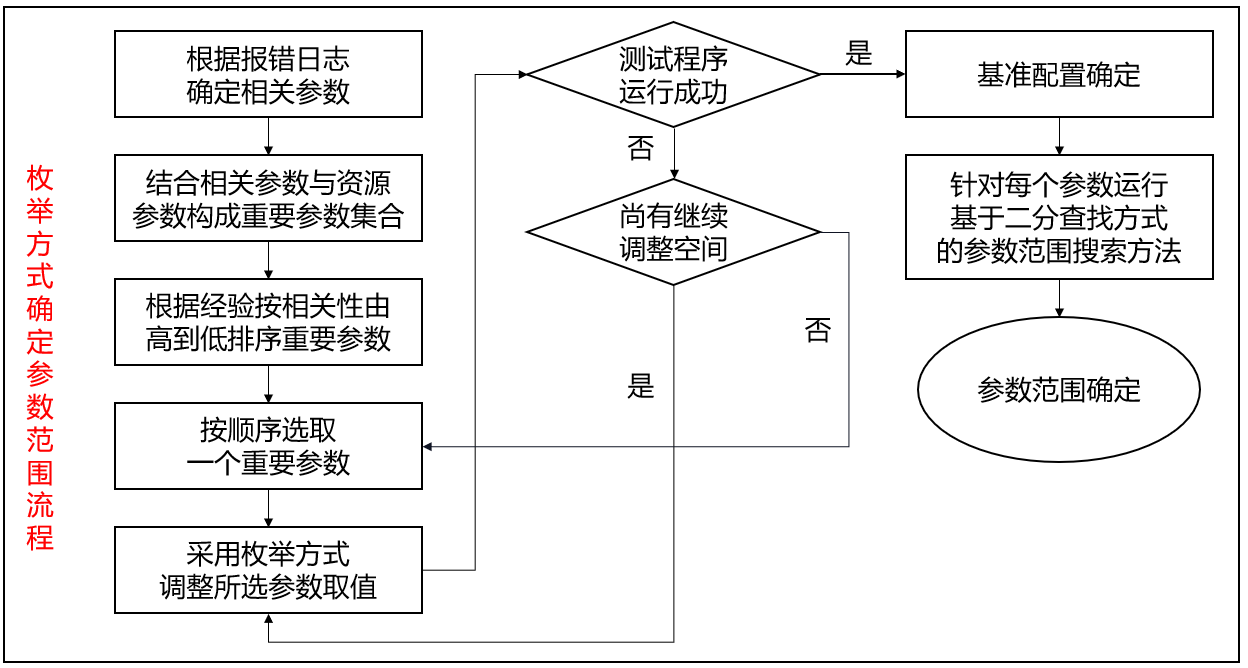


图2.7枚举法

**问题与思考**：可尝试开发自动化参数范围探测工具；可在参数范围探测时加入参数相关性考虑，以提高方法执行效率，提升最终结果的准确性。

### 2.3.2、优化方案一：搜索法

首先确定参数范围，随后使用遗传算法通过实际运行方式搜索最优配置，流程图如图2.8所示。

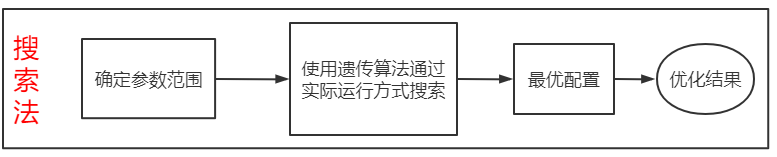


图2.8搜索法流程图

遗传算法使用细节如图2.9所示。

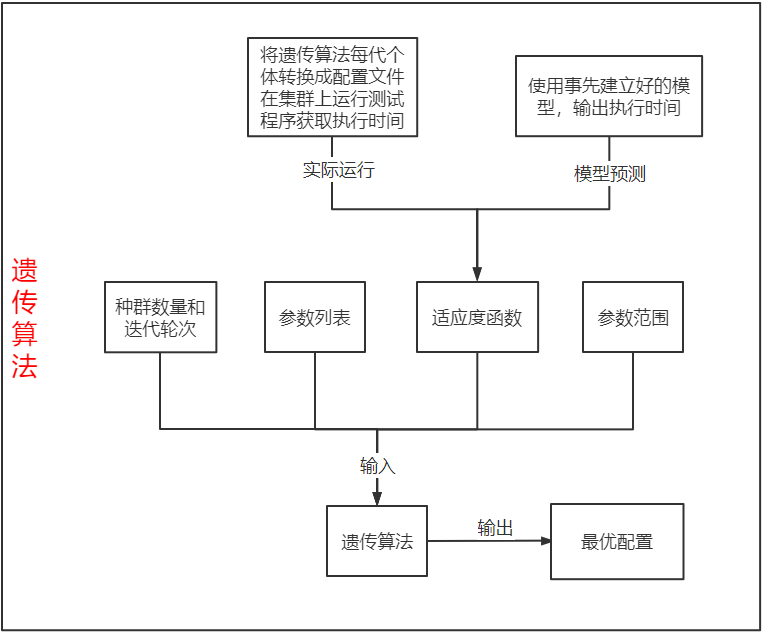


图2.9遗传算法

遗传算法的输入由四部分组成：种群数量和迭代轮次、参数列表、参数范围、以及适应度函数。其中适应度函数有两种来源，一种是通过将遗传算法运行过程中每代产生的个体转化为配置文件，在集群上运行测试程序获取执行时间后作为对应的适应度函数值，这种方式优化方案一以及后续的优化方案二中会使用到；另一种是通过事先建立好的模型，将遗传算法运行过程中每代产生的个体传入模型中，模型就会输出执行时间作为对应的适应度函数的值，这种方式后续的优化方案三中会使用到。

### 2.3.3、优化方案二：交互强度分析法

本方案分为四个阶段：**数据采集处理阶段**、**重要指标筛选阶段**、**交互强度分析阶段**、以及**遗传算法搜索阶段**，流程图如图2.10所示。

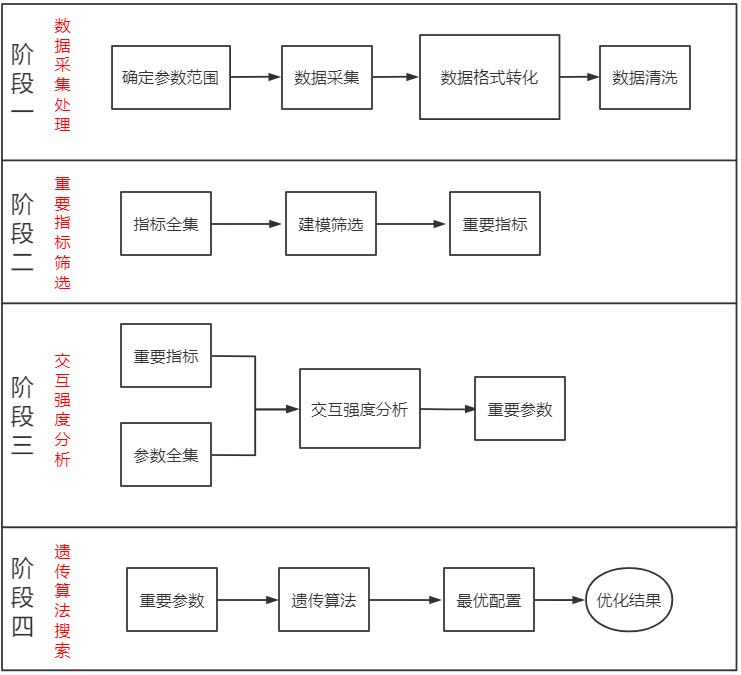


图2.10交互强度分析法

阶段一：首先进行参数范围的确定；随后在参数范围内随机生成配置文件运行测试程序进行**数据采集**，采集运行成功的测试程序的容器层，OS层，微体系结构层指标，细节如图2.11所示，采集所用的工具见图2.12，采集的部分指标见表2.14，采集的全部指标见附件:各层指标.xlsx；采集到的数据进行格式转换，**格式转换过程**将采集到的数据统一转换成CSV格式，每行代表某一秒采集到的所有指标的取值，每列代表某个指标不同时刻的取值；对于转换好的CSV格式数据进行清洗，**数据清洗过程**参考论文CounterMiner Mining Big Performance Data from Hardware Counters，处理流程如图2.13所示。



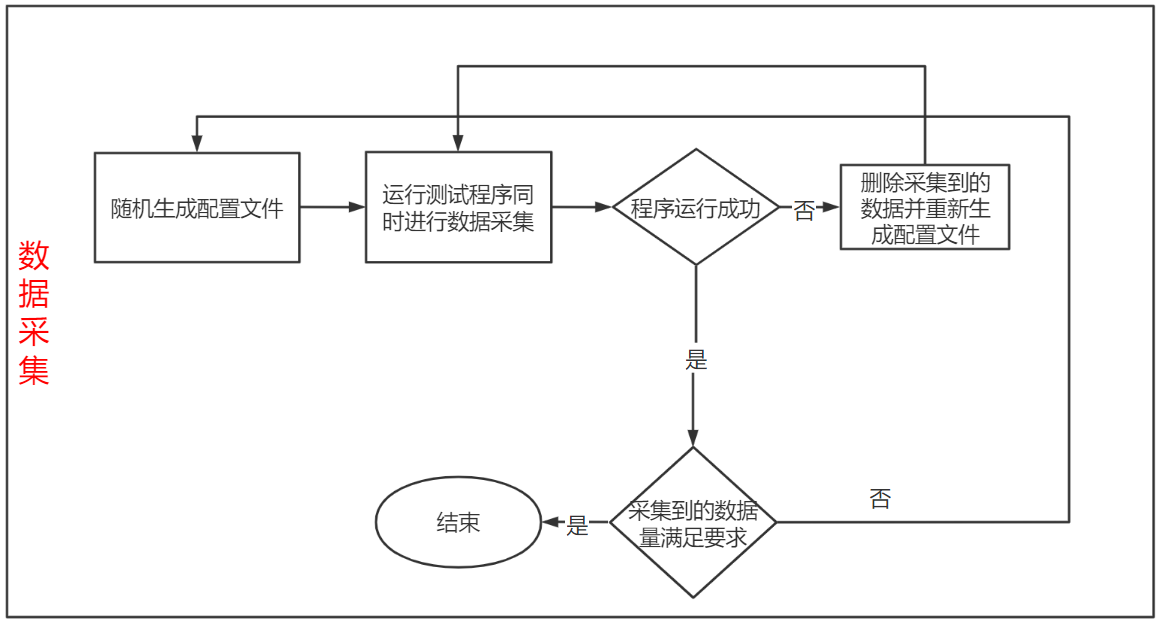


图2.11数据采集

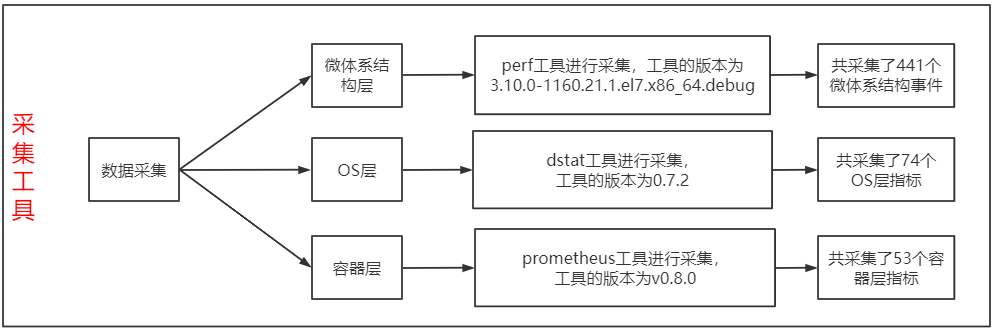


图2.12采集工具

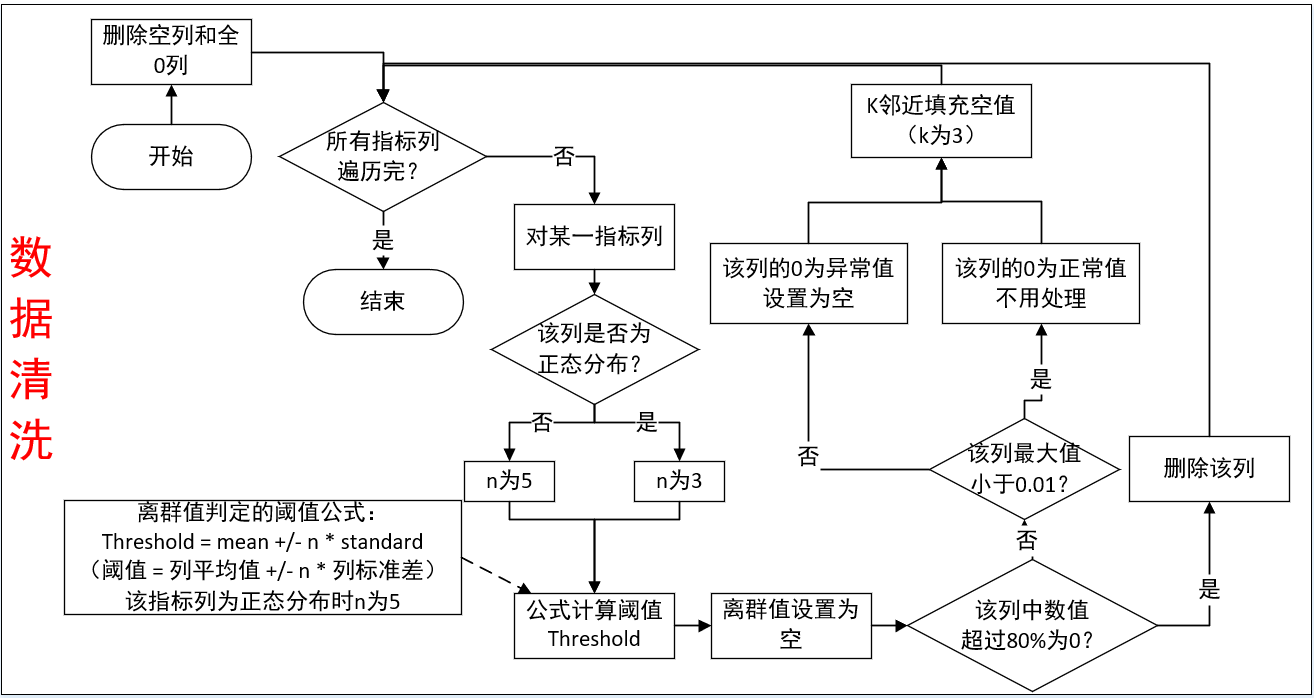


图2.13数据清洗

|  |  |  |
| --- | --- | --- |
| 微体系结构层 | OS层 | 容器层 |
| branch-instructions | load avg - 1m | kube\_pod\_container\_resource\_limits\_memory |
| branch-misses | load avg - 5m | kube\_pod\_container\_resource\_requests\_cpu |
| bus-cycles | load avg - 15m | kube\_pod\_container\_resource\_requests\_memory |
| cache-misses | total cpu usage - usr | container\_cpu\_system\_seconds\_total |
| cache-references | total cpu usage - sys | container\_cpu\_usage\_seconds\_total |

表2.14部分采集指标

阶段二：利用阶段一采集到的数据，将全部指标作为模型的特征，进行建模筛选重要指标。筛选方法为，当特征总数大于0时，使用lgb（LightGBM）进行建模，模型输入为所有指标的值，输出为IPC的值（微体系结构层能反应程序性能的指标），每次建模输出模型、特征重要性、误差、以及此次建模所使用的特征列表，删除重要性最低的特征，再进行特征数是否大于0的判断，重复以上过程。直到特征数为0时就终止，最终输出误差最低的模型以及该模型建模时使用的特征列表，具体流程图如图2.15所示。重要特征的筛选方法在后续优化方案三中也有使用。

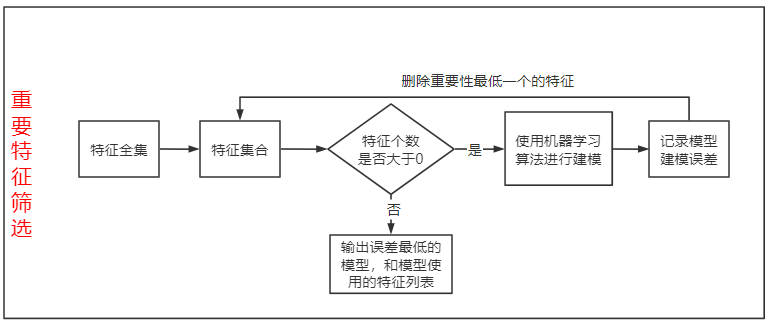


图2.15重要特征筛选

阶段三：使用lgb构建重要指标和参数预测IPC的模型M（见图2.17），重要指标和参数两两组合构成交互对，对于每一个交互对构建一组数据X（数据构建细节见图2.18），使用模型M预测出构建数据X的IPC，将（X，IPC）作为训练数据，使用线性模型（Ordinary Least Square）进行训练输出残差。计算出所有交互对的残差后结算所有交互对的交互强度（交互对D的交互强度=交互对D的残差/所有交互对残差之和），将排序靠前（排序相邻的两个交互对的交互强度比值超过1e5时前面的交互对即为排序靠前的交互对）的交互对选出来，这些交互对中参数即是筛选出来的重要参数，具体见流程图2.16。

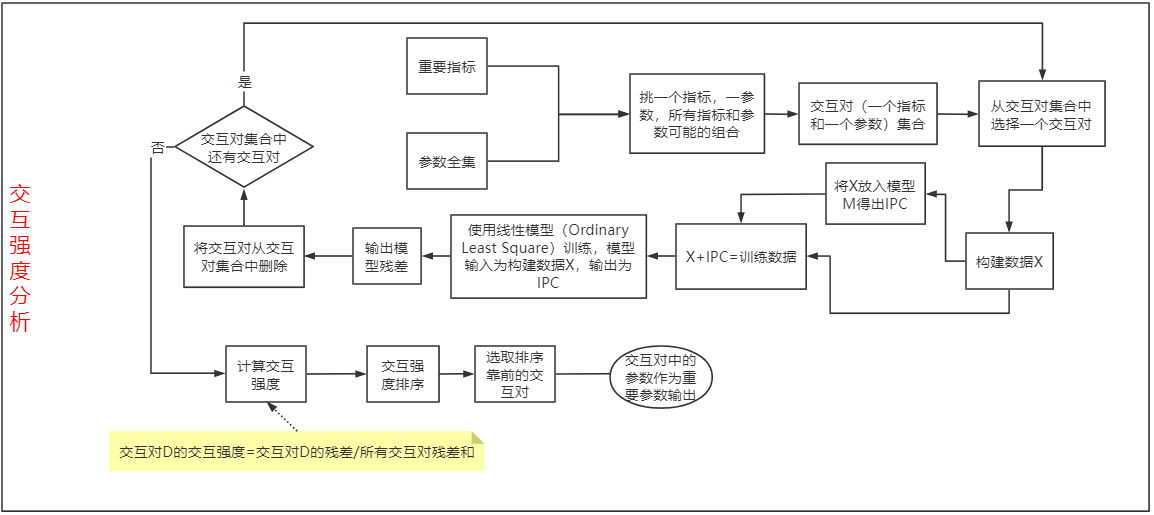


图2.16交互强度

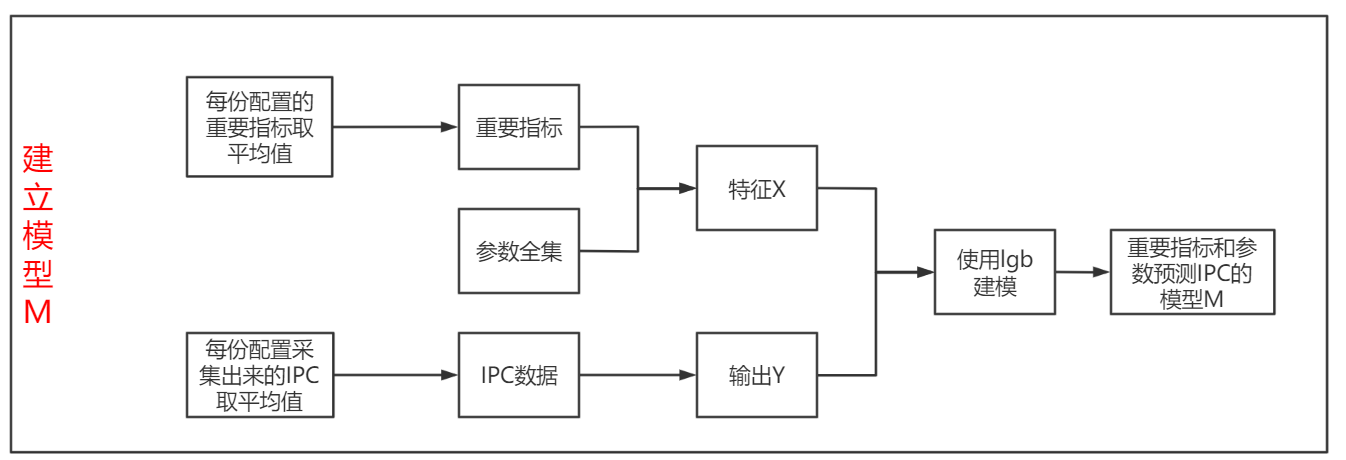


图2.17建立模型M

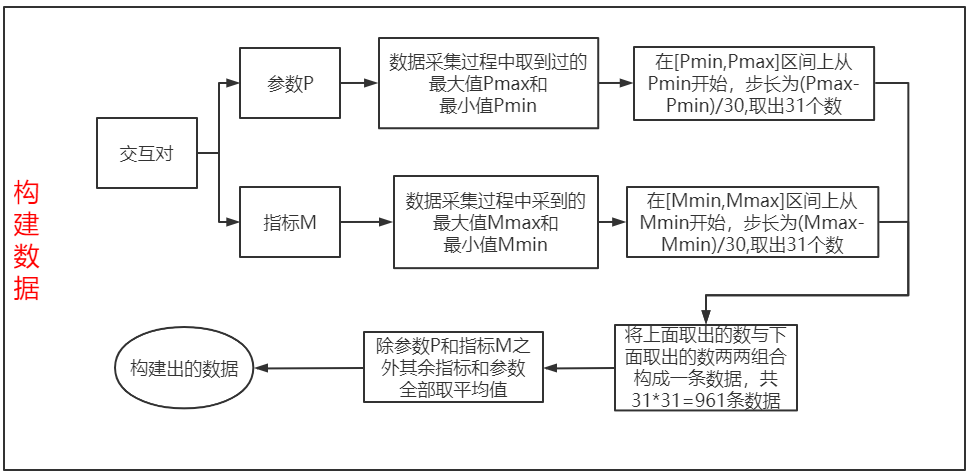


图2.18构建数据

### 2.3.4、优化方案三：迭代建模法

本方案分为三个阶段：**数据预采集阶段**、**迭代建模阶段**、以及**遗传算法搜索阶段**，流程图如图2.19所示。

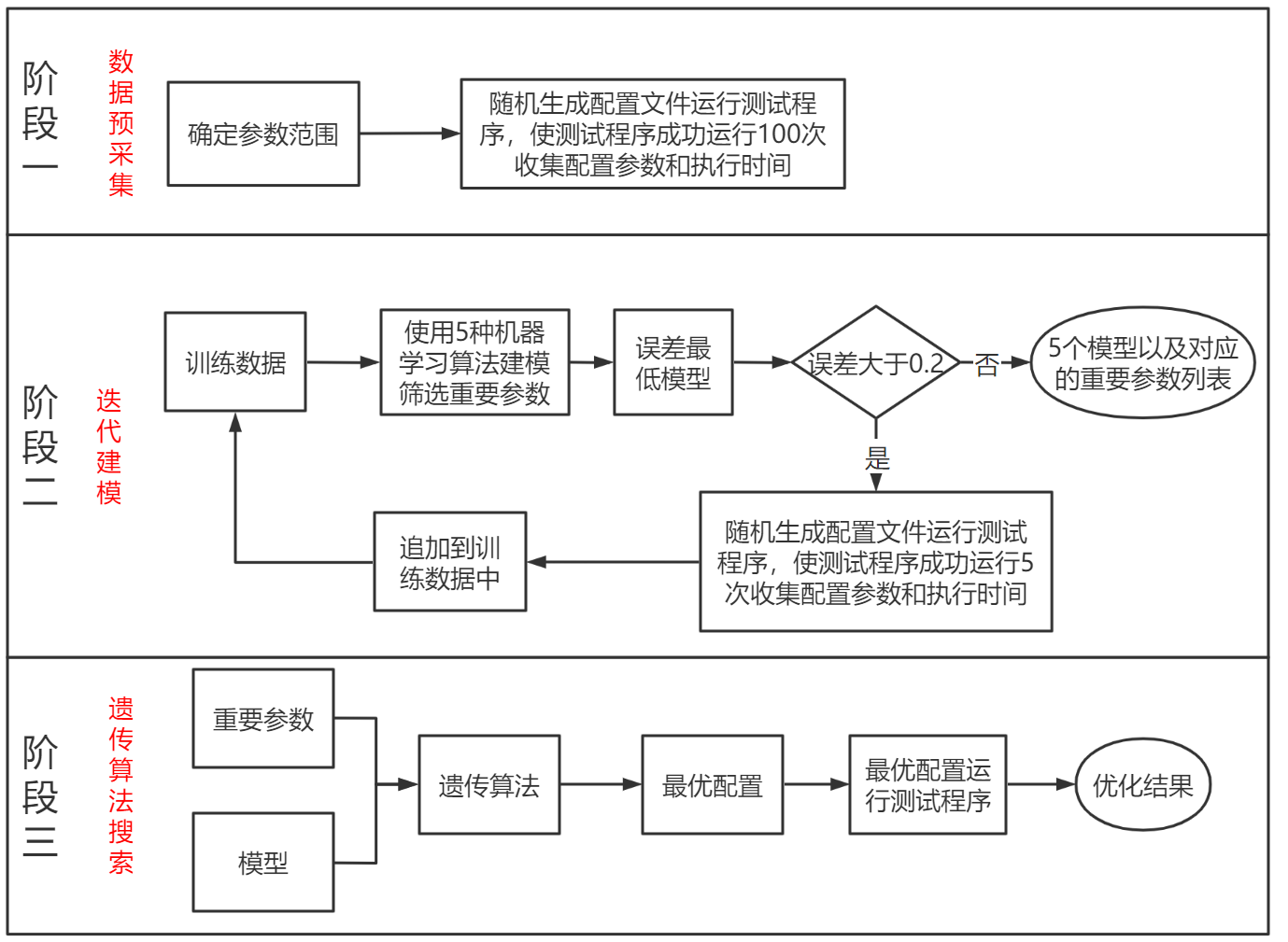


图2.19迭代建模法

阶段一：首先进行参数范围的确定；随后随机生成配置文件运行测试程序，是测试程序运行成功100次，采集运行成功测试程序的配置文件的配置参数取值情况以及程序运行时间作为后续阶段二的训练数据。

阶段二：使用5种机器学习算法（lgb：LightGBM； gdbt：Gradient Boosting Decision Tree； rf：Random Forest；ada： Adaboost；xgb：xgboost），每个算法都使用训练数据将参数取值作为输入，程序执行时间作为输出建模，通过重要特征筛选的方式（见方案二中图2.15）输出误差最低模型和模型所使用的重要特征，即输出模型和重要参数。在5个算法的结果中挑选误差最低的模型。如果该模型误差大于0.2就再增加5份训练数据（随机生成配置文件使测试程序运行成功5次，收集配置参数和执行时间），回到阶段二的开头；如果该模型误差低于0.2则输出此轮建模产生的5个模型和对应的重要参数列表，具体见图2.20。

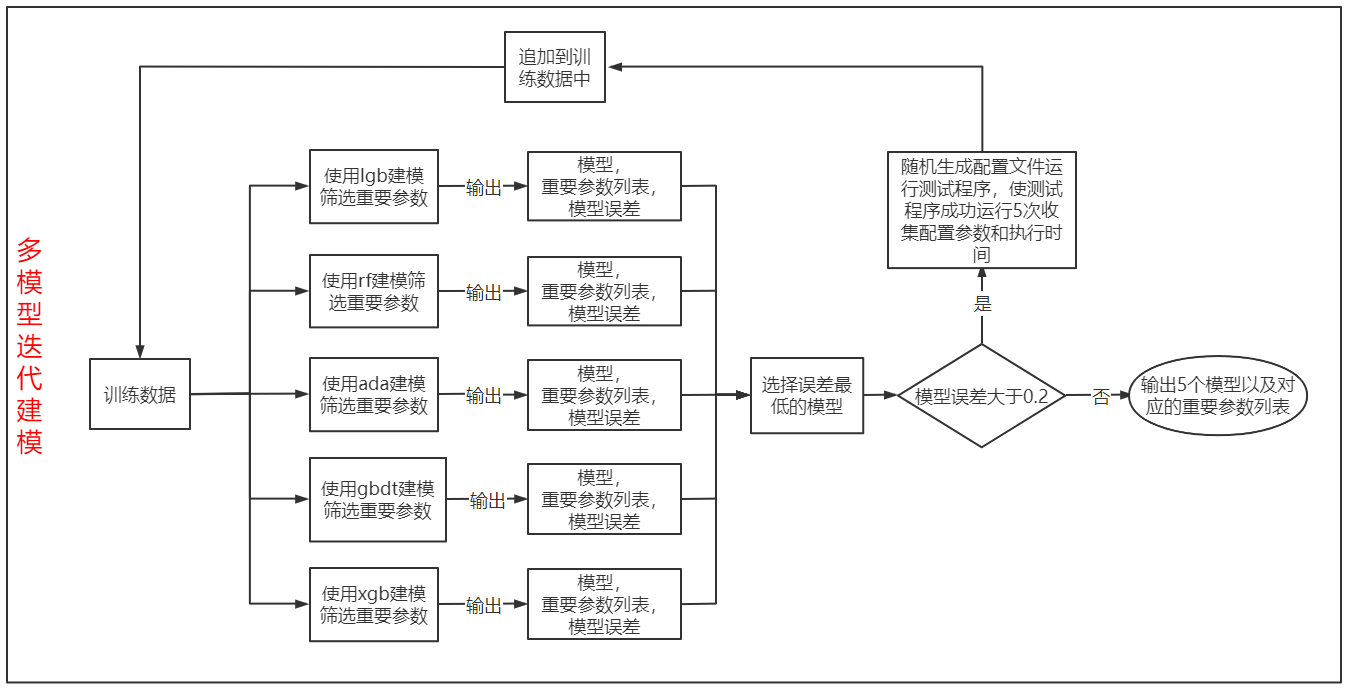


图2.20多模型迭代建模

阶段三：将5个模型和对应的重要参数列表分别放入遗传算法搜索得出最优配置，每个模型都将搜索出3份最优配置。最终将输出的15份配置文件放在机器上实际运行后观测最优结果。

### 2.3.5 三种方案对比分析

|  |  |  |  |
| --- | --- | --- | --- |
|  | 方案一 | 方案二 | 方案三 |
| **技术方案介绍** | 搜索法 | 交互强度分析法 | 迭代建模法 |
| **方案优势** | 方案执行简单，优化耗时短，有时可能会搜索到优化效果很好的配置 | 通过系统性能指标反映重要参数，可以准确找出重要参数，优化效果好 | 通过迭代建模的方式减少了样本采集，降低优化耗时。 |
| **方案劣势** | 优化结果不稳定，容易陷入局部最优，过早收敛，无法得出最优配置 | 方案执行复杂，优化耗时长 | 选择出来的重要参数列表可能不准确，导致无法得到好的优化效果 |

### 2.3.6、敏感query探测方法

敏感query探测是为了解决TPC-DS优化耗时过长，其主要流程如图2.21所示，首先随机生成配置文件执行测试程序使所有query都执行成功至少30次，收集所有query的执行时间，计算query执行时间的标准差，取标准差大于50的query作为敏感query。

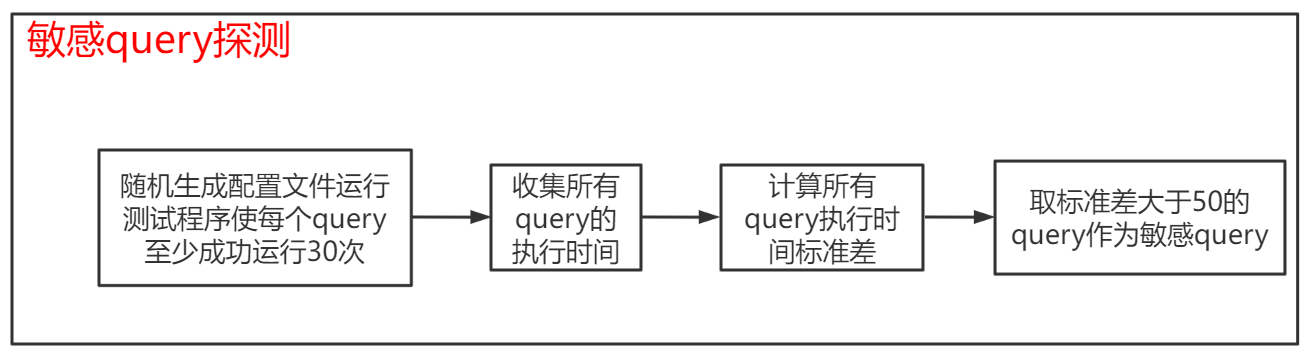


图2.21敏感query探测

**问题与思考**：此方法选出的基本上都是执行时间较长的query，可继续调研看是否有更好的选择敏感query的方法。

## 2.4、优化过程数据和结果展示

我们使用方案一对Wordcount-100G和Terasort-20G进行了优化；使用方案二对Wordcount-100G和Terasort-20G进行了优化；使用方案三对Wordcount-100G、Terasort-20G和tpcds-21G进行了优化。下面将展示优化过程数据和结果。

### 2.4.1、优化结果（方案一）

我们使用方案一对Wordcount-100G和Terasort-20G进行了优化。

#### 2.4.1.1、Wordcount-100G

优化时使用到的参数如表2.22所示，参数范围如表2.23所示

|  |  |
| --- | --- |
| spark.broadcast.blockSize | spark.files.useFetchCache |
| spark.broadcast.checksum | spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version |
| spark.broadcast.compress | spark.io.compression.codec |
| spark.cleaner.periodicGC.interval | spark.io.compression.lz4.blockSize |
| spark.cleaner.referenceTracking | spark.io.compression.snappy.blockSize |
| spark.cleaner.referenceTracking.blocking | spark.kryo.referenceTracking |
| spark.cleaner.referenceTracking.blocking.shuffle | spark.kryoserializer.buffer |
| spark.cleaner.referenceTracking.cleanCheckpoints | spark.kryoserializer.buffer.max |
| spark.default.parallelism | spark.locality.wait |
| spark.driver.cores | spark.maxRemoteBlockSizeFetchToMem |
| spark.driver.maxResultSize | spark.memory.fraction |
| spark.driver.memory | spark.memory.offHeap.enabled |
| spark.driver.memoryOverhead | spark.memory.offHeap.size |
| spark.executor.cores | spark.memory.storageFraction |
| spark.executor.heartbeatInterval | spark.memory.useLegacyMode |
| spark.executor.instances | spark.network.timeout |
| spark.executor.memory | spark.port.maxRetries |
| spark.files.fetchTimeout | spark.python.worker.memory |
| spark.files.maxPartitionBytes | spark.python.worker.reuse |
| spark.files.openCostInBytes | spark.rdd.compress |
| spark.files.overwrite | spark.reducer.maxBlocksInFlightPerAddress |
| spark.reducer.maxReqsInFlight | spark.shuffle.service.index.cache.size |
| spark.reducer.maxSizeInFlight | spark.shuffle.sort.bypassMergeThreshold |
| spark.rpc.io.backLog | spark.shuffle.spill.compress |
| spark.rpc.lookupTimeout | spark.speculation |
| spark.rpc.message.maxSize | spark.speculation.interval |
| spark.rpc.retry.wait | spark.speculation.multiplier |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | spark.speculation.quantile |
| spark.scheduler.listenerbus.eventqueue.capacity | spark.stage.maxConsecutiveAttempts |
| spark.scheduler.minRegisteredResourcesRatio | spark.storage.memoryFraction |
| spark.scheduler.mode | spark.storage.memoryMapThreshold |
| spark.scheduler.revive.interval | spark.storage.replication.proactive |
| spark.serializer | spark.storage.unrollFraction |
| spark.serializer.objectStreamReset | spark.streaming.backpressure.enabled |
| spark.shuffle.compress | spark.streaming.blockInterval |
| spark.shuffle.file.buffer | spark.streaming.receiver.writeAheadLog.enable |
| spark.shuffle.io.backLog | spark.streaming.stopGracefullyOnShutdown |
| spark.shuffle.io.maxRetries | spark.streaming.unpersist |
| spark.shuffle.io.numConnectionsPerPeer | spark.task.maxFailures |
| spark.shuffle.io.preferDirectBufs | spark.task.reaper.enabled |
| spark.shuffle.maxChunksBeingTransferred | spark.task.reaper.pollingInterval |
| spark.shuffle.memoryFraction | spark.task.reaper.threadDump |

表2.22方案一Wordcount-100G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.cores | 1-12 |
| spark.driver.maxResultSize | 1-500m |
| spark.driver.memory | 1-20g |
| spark.driver.memoryOverhead | 384M-driver.memory\*0.1 |
| spark.executor.memory | 1-16g |
| spark.python.worker.memory | 128-1000m |
| spark.python.worker.reuse | TRUE FALSE |
| spark.reducer.maxSizeInFlight | 1-500m |
| spark.reducer.maxReqsInFlight | 10-2147483647 |
| spark.reducer.maxBlocksInFlightPerAddress | 10-2147483647 |
| spark.maxRemoteBlockSizeFetchToMem | 10-2147483647 |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 1-500k |
| spark.shuffle.io.maxRetries | 2-10 |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.io.preferDirectBufs | TRUE FALSE |
| spark.shuffle.io.backLog | 8-128 |
| spark.shuffle.service.index.cache.size | 50-200m |
| spark.shuffle.maxChunksBeingTransferred | 100-9223372036854775807 |
| spark.shuffle.sort.bypassMergeThreshold | 100-1000 |
| spark.broadcast.compress | TRUE FALSE |
| spark.io.compression.codec | lz4,snappy,lzf |
| spark.io.compression.lz4.blockSize | 2-256k |
| spark.io.compression.snappy.blockSize | 2-256k |
| spark.kryo.referenceTracking | TRUE FALSE |
| spark.kryoserializer.buffer.max | 8-128m |
| spark.kryoserializer.buffer | 2-128k |
| spark.rdd.compress | TRUE FALSE |
| spark.serializer | org.apache.spark.serializer.JavaSerializer org.apache.spark.serializer.KryoSerializer |
| spark.memory.fraction | 0.5-1 |
| spark.memory.storageFraction | 0.5-1 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 10-1000m |
| spark.memory.useLegacyMode | TRUE FALSE |
| spark.storage.replication.proactive | TRUE FALSE |
| spark.cleaner.periodicGC.interval | 10-90min |
| spark.cleaner.referenceTracking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking.shuffle | TRUE FALSE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE FALSE |
| spark.broadcast.blockSize | 1-500m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.executor.cores | 1-10 |
| spark.default.parallelism | 8-500 |
| spark.executor.heartbeatInterval | 5-20s |
| spark.files.fetchTimeout | 20s-90s |
| spark.files.useFetchCache | TRUE FALSE |
| spark.files.overwrite | TRUE FALSE |
| spark.files.maxPartitionBytes | 33554432-268435456 |
| spark.files.openCostInBytes | 2097152-8388608 |
| spark.storage.memoryMapThreshold | 50-500m |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1,2 |
| spark.rpc.message.maxSize | 1-500M |
| spark.rpc.io.backLog | 32-128 |
| spark.network.timeout | 30-500s |
| spark.port.maxRetries | 4-32 |
| spark.rpc.numRetries | 2-5 |
| spark.rpc.retry.wait | 2-5 |
| spark.rpc.lookupTimeout | 30-240s |
| spark.locality.wait | 1-10s |
| spark.scheduler.minRegisteredResourcesRatio | 0.7-0.9 |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 2-50s |
| spark.scheduler.listenerbus.eventqueue.capacity | 2000-20000 |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 30-240s |
| spark.speculation | TRUE FALSE |
| spark.speculation.interval | 10-1000ms |
| spark.speculation.multiplier | 1.0-6.0 |
| spark.speculation.quantile | 0-1 |
| spark.task.maxFailures | 2-8 |
| spark.task.reaper.enabled | TRUE FALSE |
| spark.task.reaper.pollingInterval | 3-20s |
| spark.task.reaper.threadDump | TRUE FALSE |
| spark.stage.maxConsecutiveAttempts | 4-8 |
| spark.streaming.backpressure.enabled | TRUE FALSE |
| spark.streaming.blockInterval | 50-500ms |
| spark.streaming.receiver.writeAheadLog.enable | TRUE FALSE |
| spark.streaming.unpersist | TRUE FALSE |
| spark.streaming.stopGracefullyOnShutdown | TRUE FALSE |
| spark.shuffle.memoryFraction | 0-1 |
| spark.storage.memoryFraction | 0-1 |
| spark.storage.unrollFraction | 0-1 |
| spark.executor.instances | 2-14 |
| spark.shuffle.spill | TRUE FALSE |
| spark.memory.storageFraction | 0.5-1 |

表2.23方案一Wordcount-100G参数范围

优化结果：如表2.24所示，设置遗传算法种群数量30，迭代轮次10，共运行300份配置文件，优化过程数据如图2.25所示，每代最优个体如图2.26所示，最优配置的执行时间为116.824s，最优配置如表2.27所示，Wordcount-100G基准配置运行时间为1100s，优化9.4倍，优化耗时1052min。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 遗传算法输入 | | 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 种群数量 | 迭代轮次 |
| 30 | 10 | 116.824s | 全部参数使用默认值 | 1100s | 9.4 | 1052min |

表2.24方案一Wordcount-100G优化结果

图2.25方案一Wordcount-100G遗传算法所有配置执行时间

图2.26方案一Wordcount-100G遗传算法每代最优配置执行时间

|  |  |
| --- | --- |
| spark.broadcast.blockSize | 45m |
| spark.broadcast.checksum | FALSE |
| spark.broadcast.compress | TRUE |
| spark.cleaner.periodicGC.interval | 90min |
| spark.cleaner.referenceTracking | FALSE |
| spark.cleaner.referenceTracking.blocking | FALSE |
| spark.cleaner.referenceTracking.blocking.shuffle | FALSE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE |
| spark.default.parallelism | 74 |
| spark.driver.cores | 11 |
| spark.driver.maxResultSize | 369m |
| spark.driver.memory | 20g |
| spark.driver.memoryOverhead | 503M |
| spark.executor.cores | 10 |
| spark.executor.heartbeatInterval | 15s |
| spark.executor.instances | 7 |
| spark.executor.memory | 6g |
| spark.files.fetchTimeout | 58s |
| spark.files.maxPartitionBytes | 232599466 |
| spark.files.openCostInBytes | 7321642 |
| spark.files.overwrite | FALSE |
| spark.files.useFetchCache | FALSE |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 2 |
| spark.io.compression.codec | lzf |
| spark.io.compression.lz4.blockSize | 247k |
| spark.io.compression.snappy.blockSize | 195k |
| spark.kryo.referenceTracking | TRUE |
| spark.kryoserializer.buffer | 21k |
| spark.kryoserializer.buffer.max | 56m |
| spark.locality.wait | 6s |
| spark.maxRemoteBlockSizeFetchToMem | 682400139m |
| spark.memory.fraction | 0.52 |
| spark.memory.offHeap.enabled | TRUE |
| spark.memory.offHeap.size | 318m |
| spark.memory.storageFraction | 0.77 |
| spark.memory.useLegacyMode | TRUE |
| spark.network.timeout | 418s |
| spark.port.maxRetries | 17 |
| spark.python.worker.memory | 140m |
| spark.python.worker.reuse | FALSE |
| spark.rdd.compress | FALSE |
| spark.reducer.maxBlocksInFlightPerAddress | 1253580827 |
| spark.reducer.maxReqsInFlight | 802985495 |
| spark.reducer.maxSizeInFlight | 109m |
| spark.rpc.io.backLog | 41 |
| spark.rpc.lookupTimeout | 126s |
| spark.rpc.message.maxSize | 354 |
| spark.rpc.retry.wait | 3s |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 240s |
| spark.scheduler.listenerbus.eventqueue.capacity | 7151 |
| spark.scheduler.minRegisteredResourcesRatio | 0.86 |
| spark.scheduler.mode | FIFO |
| spark.scheduler.revive.interval | 11s |
| spark.serializer | org.apache.spark.serializer.JavaSerializer |
| spark.serializer.objectStreamReset | 184 |
| spark.shuffle.compress | TRUE |
| spark.shuffle.file.buffer | 2k |
| spark.shuffle.io.backLog | 32 |
| spark.shuffle.io.maxRetries | 10 |
| spark.shuffle.io.numConnectionsPerPeer | 2 |
| spark.shuffle.io.preferDirectBufs | FALSE |
| spark.shuffle.maxChunksBeingTransferred | 8.53198E+18 |
| spark.shuffle.memoryFraction | 0.82 |
| spark.shuffle.service.index.cache.size | 200m |
| spark.shuffle.sort.bypassMergeThreshold | 499 |
| spark.shuffle.spill.compress | TRUE |
| spark.speculation | TRUE |
| spark.speculation.interval | 1000ms |
| spark.speculation.multiplier | 1.69 |
| spark.speculation.quantile | 0.13 |
| spark.stage.maxConsecutiveAttempts | 6 |
| spark.storage.memoryFraction | 0.13 |
| spark.storage.memoryMapThreshold | 500m |
| spark.storage.replication.proactive | TRUE |
| spark.storage.unrollFraction | 0.65 |
| spark.streaming.backpressure.enabled | FALSE |
| spark.streaming.blockInterval | 304ms |
| spark.streaming.receiver.writeAheadLog.enable | FALSE |
| spark.streaming.stopGracefullyOnShutdown | TRUE |
| spark.streaming.unpersist | FALSE |
| spark.task.maxFailures | 7 |
| spark.task.reaper.enabled | FALSE |
| spark.task.reaper.pollingInterval | 7s |
| spark.task.reaper.threadDump | TRUE |

表2.27方案一Wordcount-100G最优配置

#### 2.4.1.2、Terasort-20G

优化时使用到的参数如表2.28所示，参数范围如表2.29所示

|  |  |
| --- | --- |
| spark.reducer.maxSizeInFlight | spark.locality.wait |
| spark.shuffle.file.buffer | spark.scheduler.revive.interval |
| spark.shuffle.memoryFraction | spark.task.maxFailures |
| spark.shuffle.sort.bypassMergeThreshold | spark.shuffle.compress |
| spark.speculation.interval | spark.memory.fraction |
| spark.speculation.multiplier | spark.speculation |
| spark.speculation.quantile | spark.broadcast.compress |
| spark.broadcast.blockSize | spark.rdd.compress |
| spark.io.compression.codec | spark.serializer |
| spark.io.compression.lz4.blockSize | spark.memory.storageFraction |
| spark.io.compression.snappy.blockSize | spark.default.parallelism |
| spark.kryo.referenceTracking | spark.memory.offHeap.enabled |
| spark.kryoserializer.buffer.max | spark.memory.offHeap.size |
| spark.kryoserializer.buffer | spark.executor.instances |
| spark.driver.cores | spark.python.worker.memory |
| spark.executor.cores | spark.python.worker.reuse |
| spark.driver.memory | spark.rpc.message.maxSize |
| spark.executor.memory | spark.driver.maxResultSize |
| spark.storage.memoryFraction | spark.driver.memoryOverhead |
| spark.storage.unrollFraction | spark.reducer.maxReqsInFlight |
| spark.network.timeout | spark.reducer.maxBlocksInFlightPerAddress |
| spark.maxRemoteBlockSizeFetchToMem | spark.storage.memoryMapThreshold |
| spark.shuffle.io.maxRetries | spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version |
| spark.shuffle.io.numConnectionsPerPeer | spark.rpc.io.backLog |
| spark.shuffle.io.preferDirectBufs | spark.port.maxRetries |
| spark.shuffle.io.backLog | spark.rpc.numRetries |
| spark.shuffle.service.index.cache.size | spark.rpc.retry.wait |
| spark.shuffle.maxChunksBeingTransferred | spark.rpc.lookupTimeout |
| spark.memory.useLegacyMode | spark.scheduler.minRegisteredResourcesRatio |
| spark.storage.replication.proactive | spark.scheduler.mode |
| spark.cleaner.periodicGC.interval | spark.scheduler.listenerbus.eventqueue.capacity |
| spark.cleaner.referenceTracking | spark.scheduler.blacklist.unschedulableTaskSetTimeout |
| spark.cleaner.referenceTracking.blocking | spark.task.reaper.enabled |
| spark.cleaner.referenceTracking.blocking.shuffle | spark.task.reaper.pollingInterval |
| spark.cleaner.referenceTracking.cleanCheckpoints | spark.task.reaper.threadDump |
| spark.broadcast.checksum | spark.stage.maxConsecutiveAttempts |
| spark.executor.heartbeatInterval | spark.streaming.backpressure.enabled |
| spark.files.fetchTimeout | spark.streaming.blockInterval |
| spark.files.useFetchCache | spark.streaming.receiver.writeAheadLog.enable |
| spark.files.overwrite | spark.streaming.unpersist |
| spark.files.maxPartitionBytes | spark.streaming.stopGracefullyOnShutdown |
| spark.files.openCostInBytes |  |

表2.28方案一Terasort-20G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.cores | 1-12 |
| spark.driver.maxResultSize | 500m-2g |
| spark.driver.memory | 1-20g |
| spark.driver.memoryOverhead | 384M-driver.memory\*0.1 |
| spark.executor.memory | 8-16g |
| spark.python.worker.memory | 128-1000m |
| spark.python.worker.reuse | TRUE FALSE |
| spark.reducer.maxSizeInFlight | 1-500m |
| spark.reducer.maxReqsInFlight | 10-2147483647 |
| spark.reducer.maxBlocksInFlightPerAddress | 10-2147483647 |
| spark.maxRemoteBlockSizeFetchToMem | 10-2147483647 |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 1-500k |
| spark.shuffle.io.maxRetries | 2-10 |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.io.preferDirectBufs | TRUE FALSE |
| spark.shuffle.io.backLog | 8-128 |
| spark.shuffle.service.index.cache.size | 50-200m |
| spark.shuffle.maxChunksBeingTransferred | 100-9223372036854775807 |
| spark.shuffle.sort.bypassMergeThreshold | 100-1000 |
| spark.broadcast.compress | TRUE FALSE |
| spark.io.compression.codec | lz4,snappy,lzf |
| spark.io.compression.lz4.blockSize | 2-256k |
| spark.io.compression.snappy.blockSize | 2-256k |
| spark.kryo.referenceTracking | TRUE FALSE |
| spark.kryoserializer.buffer.max | 8-128m |
| spark.kryoserializer.buffer | 2-128k |
| spark.rdd.compress | TRUE FALSE |
| spark.serializer | org.apache.spark.serializer.JavaSerializer org.apache.spark.serializer.KryoSerializer |
| spark.memory.fraction | 0.5-1 |
| spark.memory.storageFraction | 0.5-1 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 10-1000m |
| spark.memory.useLegacyMode | TRUE FALSE |
| spark.storage.replication.proactive | TRUE FALSE |
| spark.cleaner.periodicGC.interval | 10-90min |
| spark.cleaner.referenceTracking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking.shuffle | TRUE FALSE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE FALSE |
| spark.broadcast.blockSize | 2-128m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.executor.cores | 1-10 |
| spark.default.parallelism | 8-500 |
| spark.executor.heartbeatInterval | 5-20s |
| spark.files.fetchTimeout | 20s-90s |
| spark.files.useFetchCache | TRUE FALSE |
| spark.files.overwrite | TRUE FALSE |
| spark.files.maxPartitionBytes | 33554432-268435456 |
| spark.files.openCostInBytes | 2097152-8388608 |
| spark.storage.memoryMapThreshold | 50-500m |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1,2 |
| spark.rpc.message.maxSize | 128-500M |
| spark.rpc.io.backLog | 32-128 |
| spark.network.timeout | 120-500s |
| spark.port.maxRetries | 4-32 |
| spark.rpc.numRetries | 2-5 |
| spark.rpc.retry.wait | 2-5 |
| spark.rpc.lookupTimeout | 30-240s |
| spark.locality.wait | 1-10s |
| spark.scheduler.minRegisteredResourcesRatio | 0.7-0.9 |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 2-50s |
| spark.scheduler.listenerbus.eventqueue.capacity | 2000-20000 |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 30-240s |
| spark.speculation | TRUE FALSE |
| spark.speculation.interval | 10-1000ms |
| spark.speculation.multiplier | 2.0-6.0 |
| spark.speculation.quantile | 0-1 |
| spark.task.maxFailures | 3-8 |
| spark.task.reaper.enabled | TRUE FALSE |
| spark.task.reaper.pollingInterval | 3-20s |
| spark.task.reaper.threadDump | TRUE FALSE |
| spark.stage.maxConsecutiveAttempts | 4-8 |
| spark.streaming.backpressure.enabled | TRUE FALSE |
| spark.streaming.blockInterval | 50-500ms |
| spark.streaming.receiver.writeAheadLog.enable | TRUE FALSE |
| spark.streaming.unpersist | TRUE FALSE |
| spark.streaming.stopGracefullyOnShutdown | TRUE FALSE |
| spark.shuffle.memoryFraction | 0-1 |
| spark.storage.memoryFraction | 0-1 |
| spark.storage.unrollFraction | 0-1 |
| spark.executor.instances | 8-14 |
| spark.shuffle.spill | TRUE FALSE |
| spark.memory.storageFraction | 0.5-1 |

表2.29方案一Terasort-20G参数范围

优化结果：如表2.30所示，设置遗传算法种群数量30，迭代轮次10，共运行300份配置文件，优化过程数据如图2.31所示，每代最优个体如图2.32所示，最优配置的执行时间为169.824s，最优配置如表2.33所示，Terasort-20G基准配置运行时间为1560s，优化9.2倍，优化耗时1934min。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 遗传算法输入 | | 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 种群数量 | 迭代轮次 |
| 30 | 10 | 169.824s | spark.executor.memory设置为5g，其余参数使用默认值 | 1560s | 9.2 | 1934min |

表2.30方案一Terasort-20G优化结果

图2.31方案一Terasort-20G遗传算法所有配置执行时间

图2.32方案一Terasort-20G遗传算法每代最优配置执行时间

|  |  |
| --- | --- |
| spark.reducer.maxSizeInFlight | 181m |
| spark.shuffle.file.buffer | 174k |
| spark.shuffle.memoryFraction | 0.31 |
| spark.shuffle.sort.bypassMergeThreshold | 967 |
| spark.speculation.interval | 176ms |
| spark.speculation.multiplier | 2.47 |
| spark.speculation.quantile | 0.64 |
| spark.broadcast.blockSize | 31m |
| spark.io.compression.codec | snappy |
| spark.io.compression.lz4.blockSize | 186k |
| spark.io.compression.snappy.blockSize | 155k |
| spark.kryo.referenceTracking | TRUE |
| spark.kryoserializer.buffer.max | 119m |
| spark.kryoserializer.buffer | 64k |
| spark.driver.cores | 11 |
| spark.executor.cores | 9 |
| spark.driver.memory | 18g |
| spark.executor.memory | 13g |
| spark.storage.memoryFraction | 0.43 |
| spark.storage.unrollFraction | 0.39 |
| spark.network.timeout | 380s |
| spark.locality.wait | 10s |
| spark.scheduler.revive.interval | 14s |
| spark.task.maxFailures | 3 |
| spark.shuffle.compress | TRUE |
| spark.memory.fraction | 0.62 |
| spark.speculation | TRUE |
| spark.broadcast.compress | TRUE |
| spark.rdd.compress | TRUE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.memory.storageFraction | 0.99 |
| spark.default.parallelism | 289 |
| spark.memory.offHeap.enabled | TRUE |
| spark.memory.offHeap.size | 185m |
| spark.executor.instances | 11 |
| spark.python.worker.memory | 645m |
| spark.python.worker.reuse | TRUE |
| spark.rpc.message.maxSize | 335 |
| spark.driver.maxResultSize | 1318m |
| spark.driver.memoryOverhead | 628M |
| spark.reducer.maxReqsInFlight | 379806445 |
| spark.reducer.maxBlocksInFlightPerAddress | 1254494453 |
| spark.maxRemoteBlockSizeFetchToMem | 804128363m |
| spark.shuffle.io.maxRetries | 10 |
| spark.shuffle.io.numConnectionsPerPeer | 3 |
| spark.shuffle.io.preferDirectBufs | TRUE |
| spark.shuffle.io.backLog | 112 |
| spark.shuffle.service.index.cache.size | 200m |
| spark.shuffle.maxChunksBeingTransferred | 7.89092E+17 |
| spark.memory.useLegacyMode | TRUE |
| spark.storage.replication.proactive | TRUE |
| spark.cleaner.periodicGC.interval | 89min |
| spark.cleaner.referenceTracking | FALSE |
| spark.cleaner.referenceTracking.blocking | TRUE |
| spark.cleaner.referenceTracking.blocking.shuffle | TRUE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE |
| spark.broadcast.checksum | TRUE |
| spark.executor.heartbeatInterval | 19s |
| spark.files.fetchTimeout | 21s |
| spark.files.useFetchCache | TRUE |
| spark.files.overwrite | FALSE |
| spark.files.maxPartitionBytes | 81657244 |
| spark.files.openCostInBytes | 2650886 |
| spark.storage.memoryMapThreshold | 269m |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 2 |
| spark.rpc.io.backLog | 65 |
| spark.port.maxRetries | 26 |
| spark.rpc.numRetries | 2 |
| spark.rpc.retry.wait | 1s |
| spark.rpc.lookupTimeout | 195s |
| spark.scheduler.minRegisteredResourcesRatio | 0.7 |
| spark.scheduler.mode | FIFO |
| spark.scheduler.listenerbus.eventqueue.capacity | 12609 |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 129s |
| spark.task.reaper.enabled | FALSE |
| spark.task.reaper.pollingInterval | 3s |
| spark.task.reaper.threadDump | TRUE |
| spark.stage.maxConsecutiveAttempts | 8 |
| spark.streaming.backpressure.enabled | TRUE |
| spark.streaming.blockInterval | 377ms |
| spark.streaming.receiver.writeAheadLog.enable | FALSE |
| spark.streaming.unpersist | TRUE |
| spark.streaming.stopGracefullyOnShutdown | FALSE |

图2.33方案一中Terasort-20G最优配置

### 2.4.2、优化结果（方案二）

我们使用方案二对Wordcount-100G和Terasort-20G进行了优化。

#### 2.4.2.1、Wordcount-100G

优化时使用到的参数如表2.34所示，参数范围如表2.35所示

|  |  |
| --- | --- |
| spark.broadcast.blockSize | spark.files.useFetchCache |
| spark.broadcast.checksum | spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version |
| spark.broadcast.compress | spark.io.compression.codec |
| spark.cleaner.periodicGC.interval | spark.io.compression.lz4.blockSize |
| spark.cleaner.referenceTracking | spark.io.compression.snappy.blockSize |
| spark.cleaner.referenceTracking.blocking | spark.kryo.referenceTracking |
| spark.cleaner.referenceTracking.blocking.shuffle | spark.kryoserializer.buffer |
| spark.cleaner.referenceTracking.cleanCheckpoints | spark.kryoserializer.buffer.max |
| spark.default.parallelism | spark.locality.wait |
| spark.driver.cores | spark.maxRemoteBlockSizeFetchToMem |
| spark.driver.maxResultSize | spark.memory.fraction |
| spark.driver.memory | spark.memory.offHeap.enabled |
| spark.driver.memoryOverhead | spark.memory.offHeap.size |
| spark.executor.cores | spark.memory.storageFraction |
| spark.executor.heartbeatInterval | spark.memory.useLegacyMode |
| spark.executor.instances | spark.network.timeout |
| spark.executor.memory | spark.port.maxRetries |
| spark.files.fetchTimeout | spark.python.worker.memory |
| spark.files.maxPartitionBytes | spark.python.worker.reuse |
| spark.files.openCostInBytes | spark.rdd.compress |
| spark.files.overwrite | spark.reducer.maxBlocksInFlightPerAddress |
| spark.reducer.maxReqsInFlight | spark.shuffle.service.index.cache.size |
| spark.reducer.maxSizeInFlight | spark.shuffle.sort.bypassMergeThreshold |
| spark.rpc.io.backLog | spark.shuffle.spill.compress |
| spark.rpc.lookupTimeout | spark.speculation |
| spark.rpc.message.maxSize | spark.speculation.interval |
| spark.rpc.numRetries | spark.speculation.multiplier |
| spark.rpc.retry.wait | spark.speculation.quantile |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | spark.stage.maxConsecutiveAttempts |
| spark.scheduler.listenerbus.eventqueue.capacity | spark.storage.memoryFraction |
| spark.scheduler.minRegisteredResourcesRatio | spark.storage.memoryMapThreshold |
| spark.scheduler.mode | spark.storage.replication.proactive |
| spark.scheduler.revive.interval | spark.storage.unrollFraction |
| spark.serializer | spark.streaming.backpressure.enabled |
| spark.serializer.objectStreamReset | spark.streaming.blockInterval |
| spark.shuffle.compress | spark.streaming.receiver.writeAheadLog.enable |
| spark.shuffle.file.buffer | spark.streaming.stopGracefullyOnShutdown |
| spark.shuffle.io.backLog | spark.streaming.unpersist |
| spark.shuffle.io.maxRetries | spark.task.maxFailures |
| spark.shuffle.io.numConnectionsPerPeer | spark.task.reaper.enabled |
| spark.shuffle.io.preferDirectBufs | spark.task.reaper.pollingInterval |
| spark.shuffle.maxChunksBeingTransferred | spark.task.reaper.threadDump |
| spark.shuffle.memoryFraction |  |

表2.34方案二Wordcount-100G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.cores | 1-12 |
| spark.driver.maxResultSize | 1-500m |
| spark.driver.memory | 1-20g |
| spark.driver.memoryOverhead | 384M-driver.memory\*0.1 |
| spark.executor.memory | 1-16g |
| spark.python.worker.memory | 128-1000m |
| spark.python.worker.reuse | TRUE FALSE |
| spark.reducer.maxSizeInFlight | 1-500m |
| spark.reducer.maxReqsInFlight | 10-2147483647 |
| spark.reducer.maxBlocksInFlightPerAddress | 10-2147483647 |
| spark.maxRemoteBlockSizeFetchToMem | 10-2147483647 |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 1-500k |
| spark.shuffle.io.maxRetries | 2-10 |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.io.preferDirectBufs | TRUE FALSE |
| spark.shuffle.io.backLog | 8-128 |
| spark.shuffle.service.index.cache.size | 50-200m |
| spark.shuffle.maxChunksBeingTransferred | 100-9223372036854775807 |
| spark.shuffle.sort.bypassMergeThreshold | 100-1000 |
| spark.broadcast.compress | TRUE FALSE |
| spark.io.compression.codec | lz4,snappy,lzf |
| spark.io.compression.lz4.blockSize | 2-256k |
| spark.io.compression.snappy.blockSize | 2-256k |
| spark.kryo.referenceTracking | TRUE FALSE |
| spark.kryoserializer.buffer.max | 8-128m |
| spark.kryoserializer.buffer | 2-128k |
| spark.rdd.compress | TRUE FALSE |
| spark.serializer | org.apache.spark.serializer.JavaSerializer org.apache.spark.serializer.KryoSerializer |
| spark.memory.fraction | 0.5-1 |
| spark.memory.storageFraction | 0.5-1 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 10-1000m |
| spark.memory.useLegacyMode | TRUE FALSE |
| spark.storage.replication.proactive | TRUE FALSE |
| spark.cleaner.periodicGC.interval | 10-90min |
| spark.cleaner.referenceTracking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking.shuffle | TRUE FALSE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE FALSE |
| spark.broadcast.blockSize | 1-500m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.executor.cores | 1-10 |
| spark.default.parallelism | 8-500 |
| spark.executor.heartbeatInterval | 5-20s |
| spark.files.fetchTimeout | 20s-90s |
| spark.files.useFetchCache | TRUE FALSE |
| spark.files.overwrite | TRUE FALSE |
| spark.files.maxPartitionBytes | 33554432-268435456 |
| spark.files.openCostInBytes | 2097152-8388608 |
| spark.storage.memoryMapThreshold | 50-500m |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1,2 |
| spark.rpc.message.maxSize | 1-500M |
| spark.rpc.io.backLog | 32-128 |
| spark.network.timeout | 30-500s |
| spark.port.maxRetries | 4-32 |
| spark.rpc.numRetries | 2-5 |
| spark.rpc.retry.wait | 2-5 |
| spark.rpc.lookupTimeout | 30-240s |
| spark.locality.wait | 1-10s |
| spark.scheduler.minRegisteredResourcesRatio | 0.7-0.9 |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 2-50s |
| spark.scheduler.listenerbus.eventqueue.capacity | 2000-20000 |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 30-240s |
| spark.speculation | TRUE FALSE |
| spark.speculation.interval | 10-1000ms |
| spark.speculation.multiplier | 1.0-6.0 |
| spark.speculation.quantile | 0-1 |
| spark.task.maxFailures | 2-8 |
| spark.task.reaper.enabled | TRUE FALSE |
| spark.task.reaper.pollingInterval | 3-20s |
| spark.task.reaper.threadDump | TRUE FALSE |
| spark.stage.maxConsecutiveAttempts | 4-8 |
| spark.streaming.backpressure.enabled | TRUE FALSE |
| spark.streaming.blockInterval | 50-500ms |
| spark.streaming.receiver.writeAheadLog.enable | TRUE FALSE |
| spark.streaming.unpersist | TRUE FALSE |
| spark.streaming.stopGracefullyOnShutdown | TRUE FALSE |
| spark.shuffle.memoryFraction | 0-1 |
| spark.storage.memoryFraction | 0-1 |
| spark.storage.unrollFraction | 0-1 |
| spark.executor.instances | 2-14 |
| spark.shuffle.spill | TRUE FALSE |
| spark.memory.storageFraction | 0.5-1 |

表2.35方案二Wordcount-100G参数列表

优化结果：优化结果见表2.36所示；数据采集阶段随机生成配置文件使测试成功2000次，采集运行时数据。重要指标筛选阶段误差变化如图2.37所示，误差最低为0.08，误差最低点选择的重要指标如表2.38所示，共23个，最重要的10个指标重要性如图2.39所示，交互强度计算结果见附件：Wordcount-100G交互强度.xlsx。交互强度前十的交互对如图2.40所示，筛选出来的重要参数如表2.41所示，共43个，最终设置遗传算法种群数量30，迭代轮次10，搜索中间结果见图2.42，每代最优结果见图2.43，最优配置的执行时间为88.895s，最优配置如表2.44所示，Wordcount-100G基准配置运行时间为1100s，优化12.4倍，优化耗时11062min。



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 重要指标个数 | 重要参数个数 | 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 23 | 43 | 88.895s | 全部参数使用默认值 | 1100s | 12.4 | 11062min |

表2.36方案二Wordcount-100G优化结果

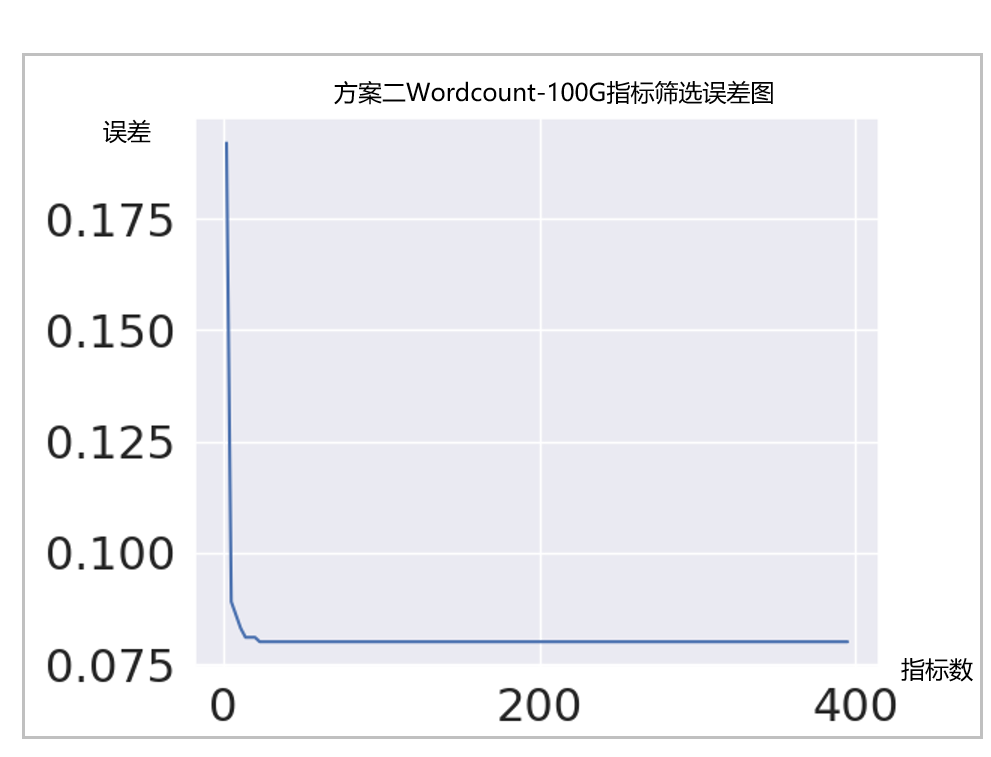


图2.37方案二Wordcount-100G指标筛选误差图

|  |  |
| --- | --- |
| L1-dcache-load-misses | topdown-total-slots |
| branch-misses | container\_start\_time\_seconds |
| msr/pperf/ | total cpu usage - idl |
| topdown-slots-retired | container\_cpu\_system\_seconds\_total |
| cache-references | topdown-recovery-bubbles |
| L1-dcache-stores | branch-instructions |
| ld\_blocks\_partial.address\_alias | memory usage - used |
| container\_spec\_cpu\_shares | exe\_activity.4\_ports\_util |
| int\_misc.recovery\_cycles | dTLB-loads |
| iTLB-load-misses | frontend\_retired.itlb\_miss |
| node-store-misses | frontend\_retired.stlb\_miss |
| unix sockets - str |  |

表2.38方案二Wordcount-100G重要指标

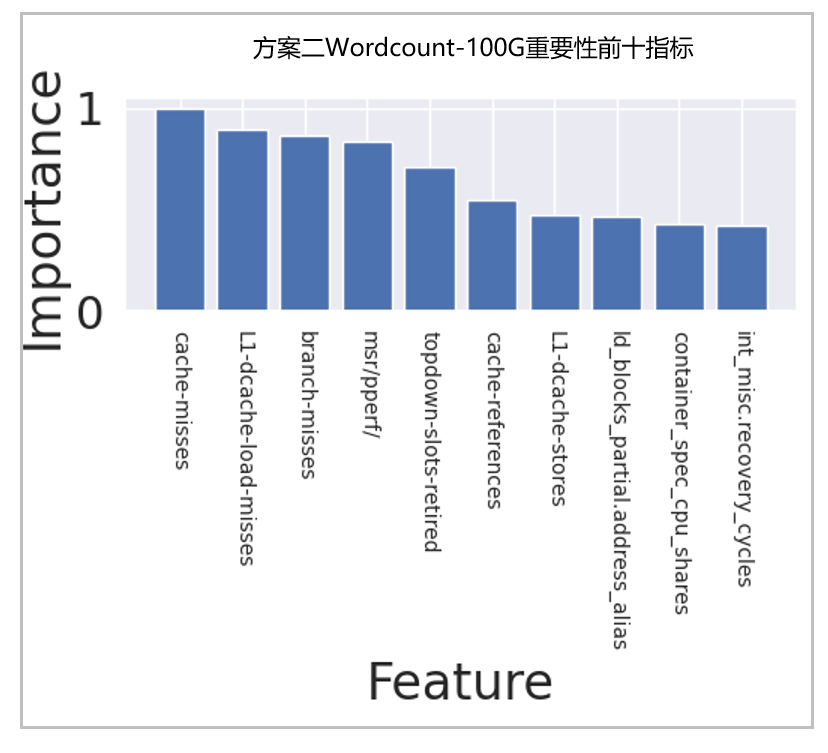


图2.39方案二Wordcount-100G重要性前十指标

图2.40方案二Wordcount-100G交互强度前十交互对

|  |  |
| --- | --- |
| spark.shuffle.service.index.cache.size | spark.scheduler.listenerbus.eventqueue.capacity |
| spark.streaming.stopGracefullyOnShutdown | spark.io.compression.snappy.blockSize |
| spark.shuffle.maxChunksBeingTransferred | spark.port.maxRetries |
| spark.speculation.multiplier | spark.streaming.backpressure.enabled |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | spark.scheduler.mode |
| spark.cleaner.referenceTracking | spark.speculation.interval |
| spark.storage.memoryMapThreshold | spark.executor.instances |
| spark.broadcast.blockSize | spark.driver.memory |
| spark.streaming.blockInterval | spark.maxRemoteBlockSizeFetchToMem |
| spark.shuffle.sort.bypassMergeThreshold | spark.rpc.message.maxSize |
| spark.serializer.objectStreamReset | spark.rpc.lookupTimeout |
| spark.executor.heartbeatInterval | spark.driver.memoryOverhead |
| spark.memory.offHeap.size | spark.driver.maxResultSize |
| spark.shuffle.io.maxRetries | spark.memory.storageFraction |
| spark.reducer.maxReqsInFlight | spark.executor.cores |
| spark.io.compression.lz4.blockSize | spark.rpc.retry.wait |
| spark.locality.wait | spark.driver.cores |
| spark.memory.fraction | spark.speculation |
| spark.executor.memory | spark.scheduler.revive.interval |
| spark.rpc.io.backLog | spark.stage.maxConsecutiveAttempts |
| spark.python.worker.memory | spark.cleaner.periodicGC.interval |
| spark.reducer.maxBlocksInFlightPerAddress |  |

表2.41方案二Wordcount-100G重要参数列表

图2.42方案二Wordcount-100G遗传算法所有配置执行时间

图2.43方案二Wordcount-100G遗传算法每代最优配置执行时间

|  |  |
| --- | --- |
| spark.shuffle.service.index.cache.size | 200m |
| spark.streaming.stopGracefullyOnShutdown | TRUE |
| spark.shuffle.maxChunksBeingTransferred | 881087284403103000 |
| spark.speculation.multiplier | 1.78 |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1 |
| spark.cleaner.referenceTracking | FALSE |
| spark.storage.memoryMapThreshold | 145m |
| spark.broadcast.blockSize | 332m |
| spark.streaming.blockInterval | 222ms |
| spark.shuffle.sort.bypassMergeThreshold | 973 |
| spark.serializer.objectStreamReset | 300 |
| spark.executor.heartbeatInterval | 20s |
| spark.memory.offHeap.size | 440m |
| spark.shuffle.io.maxRetries | 10 |
| spark.reducer.maxReqsInFlight | 1746415191 |
| spark.io.compression.lz4.blockSize | 225k |
| spark.locality.wait | 3s |
| spark.memory.fraction | 0.71 |
| spark.executor.memory | 1g |
| spark.rpc.io.backLog | 128 |
| spark.python.worker.memory | 865m |
| spark.reducer.maxBlocksInFlightPerAddress | 1986113977 |
| spark.scheduler.listenerbus.eventqueue.capacity | 20000 |
| spark.io.compression.snappy.blockSize | 135k |
| spark.port.maxRetries | 11 |
| spark.streaming.backpressure.enabled | FALSE |
| spark.scheduler.mode | FIFO |
| spark.speculation.interval | 356ms |
| spark.executor.instances | 10 |
| spark.driver.memory | 2g |
| spark.maxRemoteBlockSizeFetchToMem | 406659162m |
| spark.rpc.message.maxSize | 229 |
| spark.rpc.lookupTimeout | 133s |
| spark.driver.memoryOverhead | 1267M |
| spark.driver.maxResultSize | 377m |
| spark.memory.storageFraction | 0.94 |
| spark.executor.cores | 10 |
| spark.rpc.retry.wait | 1s |
| spark.driver.cores | 6 |
| spark.speculation | TRUE |
| spark.scheduler.revive.interval | 5s |
| spark.stage.maxConsecutiveAttempts | 4 |
| spark.cleaner.periodicGC.interval | 59min |

表2.44方案二Wordcount-100G最优配置

#### 2.4.2.2、Terasort-20G

优化时使用到的参数如表2.45所示，参数范围如表2.46所示

|  |  |
| --- | --- |
| spark.reducer.maxSizeInFlight | spark.maxRemoteBlockSizeFetchToMem |
| spark.shuffle.file.buffer | spark.shuffle.io.maxRetries |
| spark.shuffle.memoryFraction | spark.shuffle.io.numConnectionsPerPeer |
| spark.shuffle.sort.bypassMergeThreshold | spark.shuffle.io.preferDirectBufs |
| spark.speculation.interval | spark.shuffle.io.backLog |
| spark.speculation.multiplier | spark.shuffle.service.index.cache.size |
| spark.speculation.quantile | spark.shuffle.maxChunksBeingTransferred |
| spark.broadcast.blockSize | spark.memory.useLegacyMode |
| spark.io.compression.codec | spark.storage.replication.proactive |
| spark.io.compression.lz4.blockSize | spark.cleaner.periodicGC.interval |
| spark.io.compression.snappy.blockSize | spark.cleaner.referenceTracking |
| spark.kryo.referenceTracking | spark.cleaner.referenceTracking.blocking |
| spark.kryoserializer.buffer.max | spark.cleaner.referenceTracking.blocking.shuffle |
| spark.kryoserializer.buffer | spark.cleaner.referenceTracking.cleanCheckpoints |
| spark.driver.cores | spark.broadcast.checksum |
| spark.executor.cores | spark.executor.heartbeatInterval |
| spark.driver.memory | spark.files.fetchTimeout |
| spark.executor.memory | spark.files.useFetchCache |
| spark.storage.memoryFraction | spark.files.overwrite |
| spark.storage.unrollFraction | spark.files.maxPartitionBytes |
| spark.network.timeout | spark.files.openCostInBytes |
| spark.locality.wait | spark.storage.memoryMapThreshold |
| spark.scheduler.revive.interval | spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version |
| spark.task.maxFailures | spark.rpc.io.backLog |
| spark.shuffle.compress | spark.port.maxRetries |
| spark.memory.fraction | spark.rpc.numRetries |
| spark.speculation | spark.rpc.retry.wait |
| spark.broadcast.compress | spark.rpc.lookupTimeout |
| spark.rdd.compress | spark.scheduler.minRegisteredResourcesRatio |
| spark.serializer | spark.scheduler.mode |
| spark.memory.storageFraction | spark.scheduler.listenerbus.eventqueue.capacity |
| spark.default.parallelism | spark.scheduler.blacklist.unschedulableTaskSetTimeout |
| spark.memory.offHeap.enabled | spark.task.reaper.enabled |
| spark.memory.offHeap.size | spark.task.reaper.pollingInterval |
| spark.executor.instances | spark.task.reaper.threadDump |
| spark.python.worker.memory | spark.stage.maxConsecutiveAttempts |
| spark.python.worker.reuse | spark.streaming.backpressure.enabled |
| spark.rpc.message.maxSize | spark.streaming.blockInterval |
| spark.driver.maxResultSize | spark.streaming.receiver.writeAheadLog.enable |
| spark.driver.memoryOverhead | spark.streaming.unpersist |
| spark.reducer.maxReqsInFlight | spark.streaming.stopGracefullyOnShutdown |
| spark.reducer.maxBlocksInFlightPerAddress |  |

表2.45方案二Terasort-20G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.cores | 1-12 |
| spark.driver.maxResultSize | 500m-2g |
| spark.driver.memory | 1-20g |
| spark.driver.memoryOverhead | 384M-driver.memory\*0.1 |
| spark.executor.memory | 8-16g |
| spark.python.worker.memory | 128-1000m |
| spark.python.worker.reuse | TRUE FALSE |
| spark.reducer.maxSizeInFlight | 1-500m |
| spark.reducer.maxReqsInFlight | 10-2147483647 |
| spark.reducer.maxBlocksInFlightPerAddress | 10-2147483647 |
| spark.maxRemoteBlockSizeFetchToMem | 10-2147483647 |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 1-500k |
| spark.shuffle.io.maxRetries | 2-10 |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.io.preferDirectBufs | TRUE FALSE |
| spark.shuffle.io.backLog | 8-128 |
| spark.shuffle.service.index.cache.size | 50-200m |
| spark.shuffle.maxChunksBeingTransferred | 100-9223372036854775807 |
| spark.shuffle.sort.bypassMergeThreshold | 100-1000 |
| spark.broadcast.compress | TRUE FALSE |
| spark.io.compression.codec | lz4,snappy,lzf |
| spark.io.compression.lz4.blockSize | 2-256k |
| spark.io.compression.snappy.blockSize | 2-256k |
| spark.kryo.referenceTracking | TRUE FALSE |
| spark.kryoserializer.buffer.max | 8-128m |
| spark.kryoserializer.buffer | 2-128k |
| spark.rdd.compress | TRUE FALSE |
| spark.serializer | org.apache.spark.serializer.JavaSerializer org.apache.spark.serializer.KryoSerializer |
| spark.memory.fraction | 0.5-1 |
| spark.memory.storageFraction | 0.5-1 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 10-1000m |
| spark.memory.useLegacyMode | TRUE FALSE |
| spark.storage.replication.proactive | TRUE FALSE |
| spark.cleaner.periodicGC.interval | 10-90min |
| spark.cleaner.referenceTracking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking | TRUE FALSE |
| spark.cleaner.referenceTracking.blocking.shuffle | TRUE FALSE |
| spark.cleaner.referenceTracking.cleanCheckpoints | TRUE FALSE |
| spark.broadcast.blockSize | 2-128m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.executor.cores | 1-10 |
| spark.default.parallelism | 8-500 |
| spark.executor.heartbeatInterval | 5-20s |
| spark.files.fetchTimeout | 20s-90s |
| spark.files.useFetchCache | TRUE FALSE |
| spark.files.overwrite | TRUE FALSE |
| spark.files.maxPartitionBytes | 33554432-268435456 |
| spark.files.openCostInBytes | 2097152-8388608 |
| spark.storage.memoryMapThreshold | 50-500m |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1,2 |
| spark.rpc.message.maxSize | 128-500M |
| spark.rpc.io.backLog | 32-128 |
| spark.network.timeout | 120-500s |
| spark.port.maxRetries | 4-32 |
| spark.rpc.numRetries | 2-5 |
| spark.rpc.retry.wait | 2-5 |
| spark.rpc.lookupTimeout | 30-240s |
| spark.locality.wait | 1-10s |
| spark.scheduler.minRegisteredResourcesRatio | 0.7-0.9 |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 2-50s |
| spark.scheduler.listenerbus.eventqueue.capacity | 2000-20000 |
| spark.scheduler.blacklist.unschedulableTaskSetTimeout | 30-240s |
| spark.speculation | TRUE FALSE |
| spark.speculation.interval | 10-1000ms |
| spark.speculation.multiplier | 2.0-6.0 |
| spark.speculation.quantile | 0-1 |
| spark.task.maxFailures | 3-8 |
| spark.task.reaper.enabled | TRUE FALSE |
| spark.task.reaper.pollingInterval | 3-20s |
| spark.task.reaper.threadDump | TRUE FALSE |
| spark.stage.maxConsecutiveAttempts | 4-8 |
| spark.streaming.backpressure.enabled | TRUE FALSE |
| spark.streaming.blockInterval | 50-500ms |
| spark.streaming.receiver.writeAheadLog.enable | TRUE FALSE |
| spark.streaming.unpersist | TRUE FALSE |
| spark.streaming.stopGracefullyOnShutdown | TRUE FALSE |
| spark.shuffle.memoryFraction | 0-1 |
| spark.storage.memoryFraction | 0-1 |
| spark.storage.unrollFraction | 0-1 |
| spark.executor.instances | 8-14 |
| spark.shuffle.spill | TRUE FALSE |
| spark.memory.storageFraction | 0.5-1 |

表2.46方案二Terasort-20G参数范围

优化结果：优化结果如表2.47所示；数据采集阶段随机生成配置文件使测试程序成功运行2000次，采集运行时数据。重要指标筛选阶段误差变化如图2.48所示，误差最低为0.113，此处为了防止出现过多交互对，选择误差不大于0.15（任务书要求）时的最少指标数，选择的重要指标如表2.49所示，共4个，交互强度计算结果见附件：Terasort-20G交互强度.xlsx。交互强度前十的交互对如图2.50所示，筛选出来的重要参数如表2.51所示，共36个，最终设置遗传算法种群数量30，迭代轮次10，搜索中间结果见图2.52，每代最优结果见图2.53，最优配置的执行时间为204.236s，最优配置如表2.54所示，Terasort-20G基准配置运行时间为1560s，优化7.6倍，优化耗时22379min。



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 重要指标个数 | 重要参数个数 | 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 4 | 36 | 204.236s | spark.executor.memory设置为5g，其余参数使用默认值 | 1560s | 7.6 | 22379min |

表2.47方案二Terasort-20G优化结果

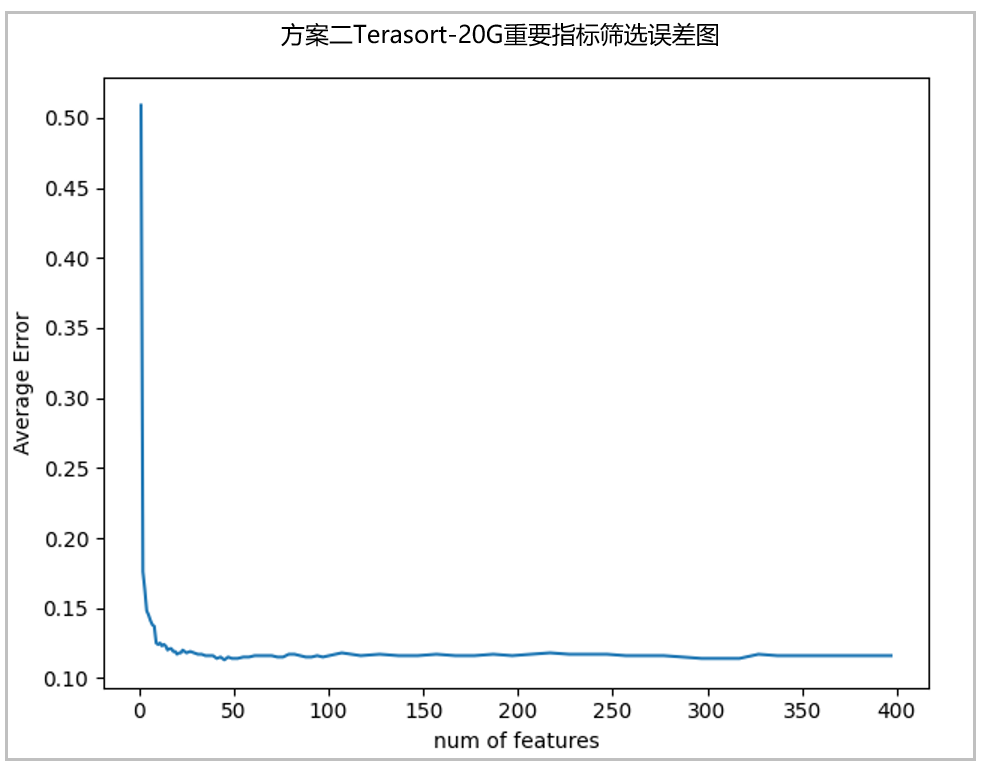


图2.48方案二Terasort-20G重要指标筛选误差图

|  |  |
| --- | --- |
| inst\_retired.any\_p | load avg - 15m |
| memory usage - used | branch-misses |

表2.49方案二Terasort-20G重要指标

图2.50方案二Terasort-20G交互强度前十交互对

|  |  |
| --- | --- |
| spark.broadcast.blockSize | spark.rpc.io.backLog |
| spark.broadcast.checksum | spark.rpc.lookupTimeout |
| spark.cleaner.periodicGC.interval | spark.rpc.retry.wait |
| spark.cleaner.referenceTracking.blocking | spark.scheduler.minRegisteredResourcesRatio |
| spark.default.parallelism | spark.scheduler.revive.interval |
| spark.executor.cores | spark.shuffle.compress |
| spark.executor.heartbeatInterval | spark.shuffle.file.buffer |
| spark.executor.instances | spark.shuffle.io.maxRetries |
| spark.executor.memory | spark.shuffle.io.numConnectionsPerPeer |
| spark.files.fetchTimeout | spark.shuffle.memoryFraction |
| spark.files.useFetchCache | spark.speculation.multiplier |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | spark.speculation.quantile |
| spark.io.compression.snappy.blockSize | spark.stage.maxConsecutiveAttempts |
| spark.kryoserializer.buffer.max | spark.storage.memoryFraction |
| spark.locality.wait | spark.storage.replication.proactive |
| spark.memory.fraction | spark.storage.unrollFraction |
| spark.memory.storageFraction | spark.task.maxFailures |
| spark.network.timeout | spark.task.reaper.threadDump |

表2.51方案二Terasort-20G重要参数列表

图2.52方案二Terasort-20G遗传算法所有配置执行时间

图2.53方案二Terasort-20G遗传算法每代最优配置执行时间

|  |  |
| --- | --- |
| spark.broadcast.blockSize | 13m |
| spark.broadcast.checksum | TRUE |
| spark.cleaner.periodicGC.interval | 90min |
| spark.cleaner.referenceTracking.blocking | TRUE |
| spark.default.parallelism | 77 |
| spark.executor.cores | 1 |
| spark.executor.heartbeatInterval | 12s |
| spark.executor.instances | 12 |
| spark.executor.memory | 16g |
| spark.files.fetchTimeout | 90s |
| spark.files.useFetchCache | TRUE |
| spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version | 1 |
| spark.io.compression.snappy.blockSize | 3k |
| spark.kryoserializer.buffer.max | 84m |
| spark.locality.wait | 4s |
| spark.memory.fraction | 0.83 |
| spark.memory.storageFraction | 0.76 |
| spark.network.timeout | 123s |
| spark.rpc.io.backLog | 35 |
| spark.rpc.lookupTimeout | 230s |
| spark.rpc.retry.wait | 3s |
| spark.scheduler.minRegisteredResourcesRatio | 0.88 |
| spark.scheduler.revive.interval | 28s |
| spark.shuffle.compress | TRUE |
| spark.shuffle.file.buffer | 315k |
| spark.shuffle.io.maxRetries | 10 |
| spark.shuffle.io.numConnectionsPerPeer | 1 |
| spark.shuffle.memoryFraction | 0.76 |
| spark.speculation.multiplier | 2.14 |
| spark.speculation.quantile | 0.29 |
| spark.stage.maxConsecutiveAttempts | 8 |
| spark.storage.memoryFraction | 0.12 |
| spark.storage.replication.proactive | TRUE |
| spark.storage.unrollFraction | 0.91 |
| spark.task.maxFailures | 4 |
| spark.task.reaper.threadDump | FALSE |

表2.54方案二Terasort-20G最优配置

### 2.4.3、优化结果（方案三）

我们使用方案三对Wordcount-100G、Terasort-20G和tpcds-21G进行了优化。

#### 2.4.3.1、Wordcount-100G

优化时使用到的参数如表2.55所示，参数范围如表2.56所示

|  |  |
| --- | --- |
| spark.memory.offHeap.enabled | spark.reducer.maxBlocksInFlightPerAddress |
| spark.memory.offHeap.size | spark.reducer.maxReqsInFlight |
| spark.executor.memory | spark.reducer.maxSizeInFlight |
| spark.executor.cores | spark.scheduler.mode |
| spark.executor.memoryOverhead | spark.scheduler.revive.interval |
| spark.executor.instances | spark.shuffle.compress |
| spark.broadcast.blockSize | spark.shuffle.file.buffer |
| spark.broadcast.checksum | spark.shuffle.io.numConnectionsPerPeer |
| spark.broadcast.compress | spark.shuffle.sort.bypassMergeThreshold |
| spark.default.parallelism | spark.storage.memoryMapThreshold |
| spark.kryoserializer.buffer | spark.driver.memory |
| spark.kryoserializer.buffer.max | spark.driver.cores |
| spark.locality.wait | spark.io.compression.codec |
| spark.maxRemoteBlockSizeFetchToMem | spark.io.compression.zstd.blockSize |
| spark.memory.fraction | spark.network.timeout |
| spark.memory.storageFraction | spark.speculation |
| spark.rdd.compress | spark.serializer |

表2.55方案三Wordcount-100G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.memory | 4G |
| spark.driver.cores | 1 |
| spark.executor.cores | 1-4 |
| spark.executor.instances | 2-8 |
| spark.executor.memory | 2-8G |
| spark.broadcast.blockSize | 2-64m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.broadcast.compress | TRUE FALSE |
| spark.default.parallelism | 200-500 |
| spark.io.compression.codec | 固定为zstd |
| spark.io.compression.zstd.blockSize | 固定为32k |
| spark.kryoserializer.buffer | 32-128k |
| spark.kryoserializer.buffer.max | 32-128m |
| spark.locality.wait | 1-10s |
| spark.maxRemoteBlockSizeFetchToMem | 1073741567-3221224702 |
| spark.memory.fraction | 0.5-0.9 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 512-1024m |
| spark.memory.storageFraction | 0.5-0.9 |
| spark.network.timeout | 120s |
| spark.rdd.compress | TRUE FALSE |
| spark.reducer.maxBlocksInFlightPerAddress | 1073741823-3221225470 |
| spark.reducer.maxReqsInFlight | 1073741823-3221225470 |
| spark.reducer.maxSizeInFlight | 24-72m |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 500ms-1000ms |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 16-48k |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.sort.bypassMergeThreshold | 100-500 |
| spark.speculation | 固定false |
| spark.storage.memoryFraction | 0.5-0.9 |
| spark.storage.memoryMapThreshold | 1-10m |

表2.56方案三Wordcount-100G参数范围

优化结果：建模搜索结果如表2.57和图2.57所示，此处使用方案三进行优化时经历了两轮阶段一和阶段二，第一轮使用全部配置参数随机生成配置使测试程序成功运行220次，使用lgb，ada，gbdt，xgb，rf建模用参数预测执行时间，筛选重要参数，最终建模误差分别为0.183, 0.246, 0.166, 0.145, 0.144。rf误差最低，筛选出来的重要参数见表2.58，共四个；第二轮使用重要参数随机生成配置使测试程序成功运行50次，使用lgb，ada，gbdt，xgb，rf建模用参数预测执行时间，此处不筛选重要参数，最终建模误差分别为0.398, 0.183, 0.007, 0.068, 0.165。将5个模型放入遗传算法，设置种群数量为400，迭代轮次为100，遗传算法结果如表2.59所示，最优配置执行时间为112.733s，对于Wordcount-100G基准配置，执行时间1100s，优化9.8倍，优化耗时1258min，见表2.60。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | lgb | ada | gdbt | xgb | rf |
| 第一轮建模误差 （220条数据） | 0.183 | 0.264 | 0.166 | 0.145 | 0.144 |
| 第二轮建模误差 （50条数据） | 0.398 | 0.183 | 0.07 | 0.068 | 0.165 |
| 遗传算法得出 最优时间（s） | - | 178.52 | 108.22 | 75.45 | 129.94 |
| 配置一实际 执行时间（s） | - | 114.653 | 115.383 | 116.945 | 124.986 |
| 配置二实际 执行时间（s） | - | 116.760 | 112.733 | 118.747 | 126.309 |
| 配置三实际 执行时间（s） | - | 118.563 | 112.819 | 117.857 | 126.792 |

表2.57方案三Wordcount-100G建模搜索结果

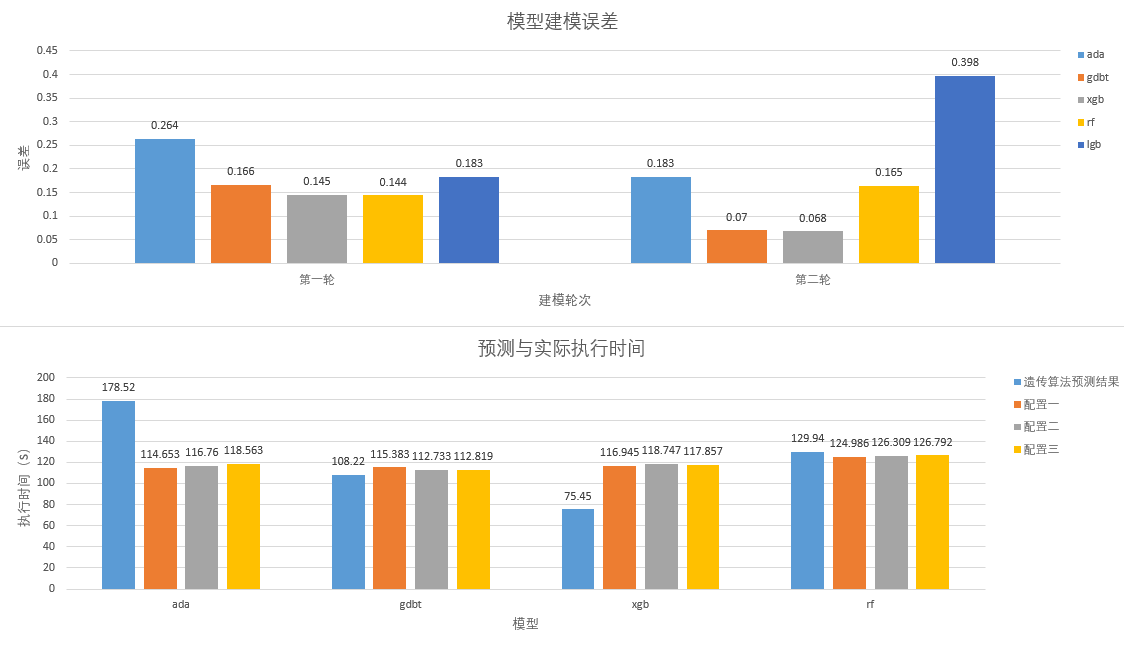


图2.57方案三Wordcount-100G建模搜索结果

|  |  |
| --- | --- |
| spark.executor.instances | spark.default.parallelism |
| spark.executor.cores | spark.memory.offHeap.size |

表2.58方案三Wordcount-100G重要参数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | 参数 | 取值 | 实际执行时间(S) | 预测时间(S) |
| 固定配置 | | spark.driver.memory | 4g | - | - |
| spark.driver.cores | 1 |
| spark.io.compression.codec | zstd |
| spark.io.compression.zstd.blockSize | 32k |
| spark.network.timeout | 120s |
| spark.speculation | FALSE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| ada | 配置一 | spark.executor.instances | 8 | 114.653 | 178.52 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 385 |
| spark.memory.offHeap.size | 597m |
| 配置二 | spark.executor.instances | 8 | 116.760 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 367 |
| spark.memory.offHeap.size | 597m |
| 配置三 | spark.executor.instances | 8 | 118.563 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 363 |
| spark.memory.offHeap.size | 577m |
| gdbt | 配置一 | spark.executor.instances | 8 | 115.383 | 108.22 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 259 |
| spark.memory.offHeap.size | 987m |
| 配置二 | spark.executor.instances | 8 | 112.733 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 233 |
| spark.memory.offHeap.size | 986m |
| 配置三 | spark.executor.instances | 8 | 112.819 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 236 |
| spark.memory.offHeap.size | 989m |
| xgb | 配置一 | spark.executor.instances | 8 | 116.945 | 75.45 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 420 |
| spark.memory.offHeap.size | 610m |
| 配置二 | spark.executor.instances | 8 | 118.747 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 418 |
| spark.memory.offHeap.size | 608m |
| 配置三 | spark.executor.instances | 8 | 117.857 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 412 |
| spark.memory.offHeap.size | 607m |
| rf | 配置一 | spark.executor.instances | 7 | 124.986 | 129.94 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 413 |
| spark.memory.offHeap.size | 620m |
| 配置二 | spark.executor.instances | 7 | 126.309 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 413 |
| spark.memory.offHeap.size | 620m |
| 配置三 | spark.executor.instances | 7 | 126.792 |
| spark.executor.cores | 4 |
| spark.default.parallelism | 415 |
| spark.memory.offHeap.size | 619m |

表2.59方案三Wordcount-100G遗传算法结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 112.733s | 全部参数使用默认值 | 1100s | 9.8 | 1258min |

表2.60方案三Wordcount-100G优化结果

#### 2.4.3.2、Terasort-20G

优化时使用到的参数如表2.61所示，参数范围如表2.62所示

|  |  |
| --- | --- |
| spark.memory.offHeap.enabled | spark.reducer.maxBlocksInFlightPerAddress |
| spark.memory.offHeap.size | spark.reducer.maxReqsInFlight |
| spark.executor.memory | spark.reducer.maxSizeInFlight |
| spark.executor.cores | spark.scheduler.mode |
| spark.executor.memoryOverhead | spark.scheduler.revive.interval |
| spark.executor.instances | spark.shuffle.compress |
| spark.broadcast.blockSize | spark.shuffle.file.buffer |
| spark.broadcast.checksum | spark.shuffle.io.numConnectionsPerPeer |
| spark.broadcast.compress | spark.shuffle.sort.bypassMergeThreshold |
| spark.default.parallelism | spark.storage.memoryMapThreshold |
| spark.kryoserializer.buffer | spark.driver.memory |
| spark.kryoserializer.buffer.max | spark.driver.cores |
| spark.locality.wait | spark.io.compression.codec |
| spark.maxRemoteBlockSizeFetchToMem | spark.io.compression.zstd.blockSize |
| spark.memory.fraction | spark.network.timeout |
| spark.memory.storageFraction | spark.speculation |
| spark.rdd.compress | spark.serializer |

表2.61方案三Terasort-20G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.driver.memory | 4G |
| spark.driver.cores | 1 |
| spark.executor.cores | 2-4 |
| spark.executor.instances | 8-16 |
| spark.executor.memory | 8-16G |
| spark.broadcast.blockSize | 2-64m |
| spark.broadcast.checksum | TRUE FALSE |
| spark.broadcast.compress | TRUE FALSE |
| spark.default.parallelism | 200-500 |
| spark.io.compression.codec | 固定为zstd |
| spark.io.compression.zstd.blockSize | 固定为32k |
| spark.kryoserializer.buffer | 32-128k |
| spark.kryoserializer.buffer.max | 32-128m |
| spark.locality.wait | 1-10s |
| spark.maxRemoteBlockSizeFetchToMem | 1073741567-3221224702 |
| spark.memory.fraction | 0.5-0.9 |
| spark.memory.offHeap.enabled | TRUE FALSE |
| spark.memory.offHeap.size | 512-1024m |
| spark.memory.storageFraction | 0.5-0.9 |
| spark.network.timeout | 120s |
| spark.rdd.compress | TRUE FALSE |
| spark.reducer.maxBlocksInFlightPerAddress | 1073741823-3221225470 |
| spark.reducer.maxReqsInFlight | 1073741823-3221225470 |
| spark.reducer.maxSizeInFlight | 24-72m |
| spark.scheduler.mode | FAIR FIFO |
| spark.scheduler.revive.interval | 500ms-1000ms |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.shuffle.compress | TRUE FALSE |
| spark.shuffle.file.buffer | 16-48k |
| spark.shuffle.io.numConnectionsPerPeer | 1-5 |
| spark.shuffle.sort.bypassMergeThreshold | 100-500 |
| spark.speculation | 固定false |
| spark.storage.memoryFraction | 0.5-0.9 |
| spark.storage.memoryMapThreshold | 1-10m |

表2.62方案三Terasort-20G参数范围

优化结果：建模搜索结果如表2.63和图2.63所示，使用全部配置参数随机生成配置使测试程序成功运行200次，使用gbdt， rf，xgb建模用参数预测执行时间，筛选重要参数，最终建模误差分别为0.211, 0.213, 0.203。各模型筛选出来的重要参数见表2.64。将3个模型放入遗传算法，设置种群数量为400，迭代轮次为100，遗传算法结果如表2.65所示，最优配置执行时间为192.389s，对于Terasort-20G基准配置，执行时间1560s，优化8.1倍，优化耗时1845min，见表2.66。

|  |  |  |  |
| --- | --- | --- | --- |
|  | gdbt | rf | xgb |
| 模型误差 | 0.211 | 0.213 | 0.203 |
| 重要参数个数 | 19 | 23 | 27 |
| 遗传算法得出  最优时间 （s） | 503.24 | 241.4 | 251.5 |
| 配置一实际  执行时间（s） | 293.873 | 207.021 | 285.432 |
| 配置二实际  执行时间（s） | 475.877 | 196.588 | 284.885 |
| 配置三实际  执行时间（s） | 238.184 | 192.389 | 236.103 |

表2.63方案三Terasort-20G建模搜索结果

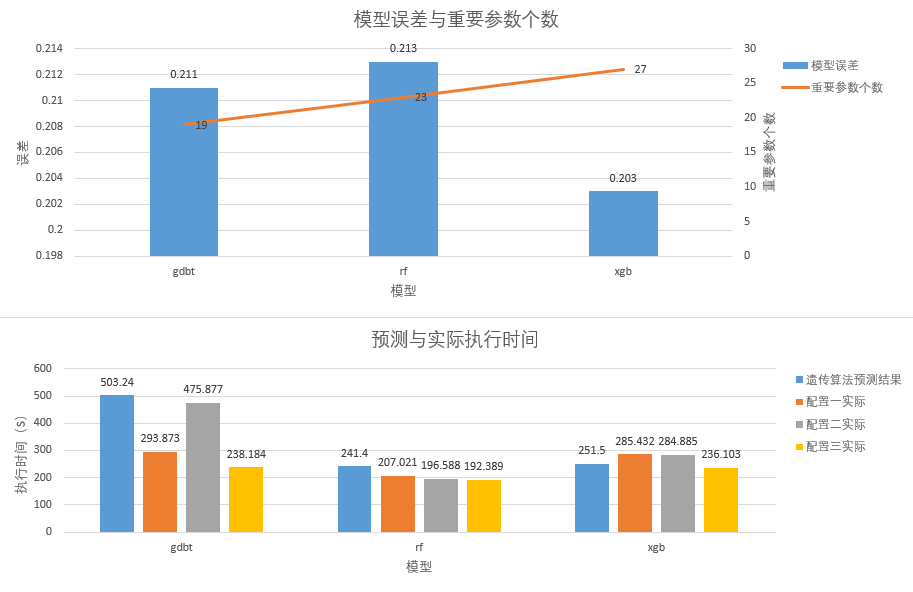


图2.63方案三Terasort-20G建模搜索结果

|  |  |
| --- | --- |
| 模型 | 参数 |
| xgb | spark.executor.cores |
| spark.executor.instances |
| spark.executor.memory |
| spark.broadcast.blockSize |
| spark.broadcast.checksum |
| spark.broadcast.compress |
| spark.default.parallelism |
| spark.kryoserializer.buffer |
| spark.kryoserializer.buffer.max |
| spark.locality.wait |
| spark.maxRemoteBlockSizeFetchToMem |
| spark.memory.fraction |
| spark.memory.offHeap.enabled |
| spark.memory.offHeap.size |
| spark.memory.storageFraction |
| spark.rdd.compress |
| spark.reducer.maxBlocksInFlightPerAddress |
| spark.reducer.maxReqsInFlight |
| spark.reducer.maxSizeInFlight |
| spark.scheduler.mode |
| spark.scheduler.revive.interval |
| spark.shuffle.compress |
| spark.shuffle.file.buffer |
| spark.shuffle.io.numConnectionsPerPeer |
| spark.shuffle.sort.bypassMergeThreshold |
| spark.storage.memoryFraction |
| spark.storage.memoryMapThreshold |
| gdbt | spark.executor.memory |
| spark.shuffle.compress |
| spark.reducer.maxBlocksInFlightPerAddress |
| spark.memory.offHeap.size |
| spark.maxRemoteBlockSizeFetchToMem |
| spark.memory.fraction |
| spark.broadcast.blockSize |
| spark.reducer.maxReqsInFlight |
| spark.scheduler.revive.interval |
| spark.executor.cores |
| spark.storage.memoryFraction |
| spark.reducer.maxSizeInFlight |
| spark.kryoserializer.buffer |
| spark.default.parallelism |
| spark.kryoserializer.buffer.max |
| spark.broadcast.compress |
| spark.executor.instances |
| spark.storage.memoryMapThreshold |
| spark.locality.wait |
| rf | spark.executor.memory |
| spark.broadcast.blockSize |
| spark.shuffle.compress |
| spark.maxRemoteBlockSizeFetchToMem |
| spark.scheduler.revive.interval |
| spark.memory.offHeap.size |
| spark.memory.fraction |
| spark.reducer.maxBlocksInFlightPerAddress |
| spark.reducer.maxSizeInFlight |
| spark.storage.memoryFraction |
| spark.default.parallelism |
| spark.kryoserializer.buffer.max |
| spark.memory.storageFraction |
| spark.kryoserializer.buffer |
| spark.reducer.maxReqsInFlight |
| spark.shuffle.file.buffer |
| spark.shuffle.sort.bypassMergeThreshold |
| spark.executor.instances |
| spark.storage.memoryMapThreshold |
| spark.executor.cores |
| spark.locality.wait |
| spark.shuffle.io.numConnectionsPerPeer |
| spark.broadcast.compress |

表2.64方案三Terasort-20G重要参数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | 参数 | 取值 | 执行时间(S) | 预测时间(S) |
| 固定配置 | | spark.driver.memory | 4g | - | - |
| spark.driver.cores | 1 |
| spark.io.compression.codec | zstd |
| spark.io.compression.zstd.blockSize | 32k |
| spark.network.timeout | 120s |
| spark.speculation | FALSE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| xgb | 配置一 | spark.executor.cores | 3 | 285.432 | 251.5 |
| spark.executor.instances | 16 |
| spark.executor.memory | 16g |
| spark.broadcast.blockSize | 3m |
| spark.broadcast.checksum | FALSE |
| spark.broadcast.compress | FALSE |
| spark.default.parallelism | 500 |
| spark.kryoserializer.buffer | 66k |
| spark.kryoserializer.buffer.max | 90m |
| spark.locality.wait | 10s |
| spark.maxRemoteBlockSizeFetchToMem | 2067637104 |
| spark.memory.fraction | 0.51 |
| spark.memory.offHeap.enabled | FALSE |
| spark.memory.offHeap.size | 1024m |
| spark.memory.storageFraction | 0.58 |
| spark.rdd.compress | TRUE |
| spark.reducer.maxBlocksInFlightPerAddress | 2147483647 |
| spark.reducer.maxReqsInFlight | 1166791289 |
| spark.reducer.maxSizeInFlight | 62m |
| spark.scheduler.mode | FIFO |
| spark.scheduler.revive.interval | 625ms |
| spark.shuffle.compress | TRUE |
| spark.shuffle.file.buffer | 48k |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.shuffle.sort.bypassMergeThreshold | 500 |
| spark.storage.memoryFraction | 0.64 |
| spark.storage.memoryMapThreshold | 10m |
| 配置二 | spark.executor.cores | 3 | 284.885 |
| spark.executor.instances | 16 |
| spark.executor.memory | 8g |
| spark.broadcast.blockSize | 3m |
| spark.broadcast.checksum | false |
| spark.broadcast.compress | true |
| spark.default.parallelism | 316 |
| spark.kryoserializer.buffer | 81k |
| spark.kryoserializer.buffer.max | 96m |
| spark.locality.wait | 1s |
| spark.maxRemoteBlockSizeFetchToMem | 1550419712 |
| spark.memory.fraction | 0.62 |
| spark.memory.offHeap.enabled | false |
| spark.memory.offHeap.size | 850m |
| spark.memory.storageFraction | 0.8 |
| spark.rdd.compress | true |
| spark.reducer.maxBlocksInFlightPerAddress | 1915040961 |
| spark.reducer.maxReqsInFlight | 2147483647 |
| spark.reducer.maxSizeInFlight | 37m |
| spark.scheduler.mode | FIFO |
| spark.scheduler.revive.interval | 633ms |
| spark.shuffle.compress | true |
| spark.shuffle.file.buffer | 40k |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.shuffle.sort.bypassMergeThreshold | 130 |
| spark.storage.memoryFraction | 0.61 |
| spark.storage.memoryMapThreshold | 6m |
| 配置三 | spark.executor.cores | 3 | 236.103 |
| spark.executor.instances | 16 |
| spark.executor.memory | 16g |
| spark.broadcast.blockSize | 2m |
| spark.broadcast.checksum | true |
| spark.broadcast.compress | true |
| spark.default.parallelism | 237 |
| spark.kryoserializer.buffer | 80k |
| spark.kryoserializer.buffer.max | 69m |
| spark.locality.wait | 10s |
| spark.maxRemoteBlockSizeFetchToMem | 1647902681 |
| spark.memory.fraction | 0.89 |
| spark.memory.offHeap.enabled | true |
| spark.memory.offHeap.size | 1024m |
| spark.memory.storageFraction | 0.58 |
| spark.rdd.compress | true |
| spark.reducer.maxBlocksInFlightPerAddress | 1800998957 |
| spark.reducer.maxReqsInFlight | 2147483647 |
| spark.reducer.maxSizeInFlight | 42m |
| spark.scheduler.mode | FIFO |
| spark.scheduler.revive.interval | 556ms |
| spark.shuffle.compress | false |
| spark.shuffle.file.buffer | 48k |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.shuffle.sort.bypassMergeThreshold | 500 |
| spark.storage.memoryFraction | 0.77 |
| spark.storage.memoryMapThreshold | 10m |
| gdbt | 配置一 | spark.executor.memory | 16g | 293.873 | 503.24 |
| spark.shuffle.compress | true |
| spark.reducer.maxBlocksInFlightPerAddress | 1946508482 |
| spark.memory.offHeap.size | 831m |
| spark.maxRemoteBlockSizeFetchToMem | 1129242525 |
| spark.memory.fraction | 0.58 |
| spark.broadcast.blockSize | 59m |
| spark.reducer.maxReqsInFlight | 2147483647 |
| spark.scheduler.revive.interval | 863ms |
| spark.executor.cores | 4 |
| spark.storage.memoryFraction | 0.55 |
| spark.reducer.maxSizeInFlight | 30m |
| spark.kryoserializer.buffer | 35k |
| spark.default.parallelism | 500 |
| spark.kryoserializer.buffer.max | 39m |
| spark.broadcast.compress | true |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 10m |
| spark.locality.wait | 1s |
| 配置二 | spark.executor.memory | 16g | 475.877 |
| spark.shuffle.compress | true |
| spark.reducer.maxBlocksInFlightPerAddress | 2011180414 |
| spark.memory.offHeap.size | 803m |
| spark.maxRemoteBlockSizeFetchToMem | 1087511463 |
| spark.memory.fraction | 0.88 |
| spark.broadcast.blockSize | 3m |
| spark.reducer.maxReqsInFlight | 1239759369 |
| spark.scheduler.revive.interval | 928ms |
| spark.executor.cores | 4 |
| spark.storage.memoryFraction | 0.65 |
| spark.reducer.maxSizeInFlight | 53m |
| spark.kryoserializer.buffer | 100k |
| spark.default.parallelism | 486 |
| spark.kryoserializer.buffer.max | 125m |
| spark.broadcast.compress | true |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 10m |
| spark.locality.wait | 1s |
| 配置三 | spark.executor.memory | 16g | 238.184 |
| spark.shuffle.compress | true |
| spark.reducer.maxBlocksInFlightPerAddress | 1844381436 |
| spark.memory.offHeap.size | 831m |
| spark.maxRemoteBlockSizeFetchToMem | 1101143524 |
| spark.memory.fraction | 0.71 |
| spark.broadcast.blockSize | 43m |
| spark.reducer.maxReqsInFlight | 1501720648 |
| spark.scheduler.revive.interval | 958ms |
| spark.executor.cores | 2 |
| spark.storage.memoryFraction | 0.73 |
| spark.reducer.maxSizeInFlight | 52m |
| spark.kryoserializer.buffer | 47k |
| spark.default.parallelism | 500 |
| spark.kryoserializer.buffer.max | 91m |
| spark.broadcast.compress | true |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 10m |
| spark.locality.wait | 1s |
| rf | 配置一 | spark.executor.memory | 13g | 207.021 | 241.4 |
| spark.broadcast.blockSize | 27m |
| spark.shuffle.compress | true |
| spark.maxRemoteBlockSizeFetchToMem | 1730777380 |
| spark.scheduler.revive.interval | 645ms |
| spark.memory.offHeap.size | 517m |
| spark.memory.fraction | 0.67 |
| spark.reducer.maxBlocksInFlightPerAddress | 1199199603 |
| spark.reducer.maxSizeInFlight | 57m |
| spark.storage.memoryFraction | 0.81 |
| spark.default.parallelism | 231 |
| spark.kryoserializer.buffer.max | 103m |
| spark.memory.storageFraction | 0.63 |
| spark.kryoserializer.buffer | 81k |
| spark.reducer.maxReqsInFlight | 2147483647 |
| spark.shuffle.file.buffer | 32k |
| spark.shuffle.sort.bypassMergeThreshold | 365 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 3m |
| spark.executor.cores | 4 |
| spark.locality.wait | 8s |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.broadcast.compress | false |
| 配置二 | spark.executor.memory | 13g | 196.588 |
| spark.broadcast.blockSize | 27m |
| spark.shuffle.compress | true |
| spark.maxRemoteBlockSizeFetchToMem | 1731753167 |
| spark.scheduler.revive.interval | 643ms |
| spark.memory.offHeap.size | 525m |
| spark.memory.fraction | 0.58 |
| spark.reducer.maxBlocksInFlightPerAddress | 1660279313 |
| spark.reducer.maxSizeInFlight | 58m |
| spark.storage.memoryFraction | 0.82 |
| spark.default.parallelism | 231 |
| spark.kryoserializer.buffer.max | 103m |
| spark.memory.storageFraction | 0.63 |
| spark.kryoserializer.buffer | 128k |
| spark.reducer.maxReqsInFlight | 1740225551 |
| spark.shuffle.file.buffer | 34k |
| spark.shuffle.sort.bypassMergeThreshold | 101 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 3m |
| spark.executor.cores | 4 |
| spark.locality.wait | 8s |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.broadcast.compress | true |
| 配置三 | spark.executor.memory | 13g | 192.389 |
| spark.broadcast.blockSize | 27m |
| spark.shuffle.compress | true |
| spark.maxRemoteBlockSizeFetchToMem | 1726370208 |
| spark.scheduler.revive.interval | 644ms |
| spark.memory.offHeap.size | 515m |
| spark.memory.fraction | 0.9 |
| spark.reducer.maxBlocksInFlightPerAddress | 2147483647 |
| spark.reducer.maxSizeInFlight | 57m |
| spark.storage.memoryFraction | 0.81 |
| spark.default.parallelism | 232 |
| spark.kryoserializer.buffer.max | 103m |
| spark.memory.storageFraction | 0.61 |
| spark.kryoserializer.buffer | 60k |
| spark.reducer.maxReqsInFlight | 1410158684 |
| spark.shuffle.file.buffer | 33k |
| spark.shuffle.sort.bypassMergeThreshold | 330 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 3m |
| spark.executor.cores | 4 |
| spark.locality.wait | 8s |
| spark.shuffle.io.numConnectionsPerPeer | 5 |
| spark.broadcast.compress | true |

表2.65方案三Terasort-20G遗传算法结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 192.389s | spark.executor.memory设置为5g，其余参数使用默认值 | 1560s | 8.1 | 1845min |

表2.66方案三Terasort-20G优化结果

#### 2.4.3.2、tpcds-21G

优化时使用到的参数如表2.67所示，参数范围如表2.68所示

|  |  |
| --- | --- |
| spark.memory.offHeap.enabled | spark.memory.storageFraction |
| spark.memory.offHeap.size | spark.rdd.compress |
| spark.executor.memory | spark.reducer.maxBlocksInFlightPerAddress |
| spark.executor.cores | spark.reducer.maxReqsInFlight |
| spark.executor.memoryOverhead | spark.reducer.maxSizeInFlight |
| spark.executor.instances | spark.scheduler.revive.interval |
| spark.broadcast.blockSize | spark.shuffle.file.buffer |
| spark.broadcast.checksum | spark.shuffle.io.numConnectionsPerPeer |
| spark.broadcast.compress | spark.shuffle.sort.bypassMergeThreshold |
| spark.default.parallelism | spark.storage.memoryMapThreshold |
| spark.kryoserializer.buffer | spark.sql.shuffle.partitions |
| spark.kryoserializer.buffer.max | spark.sql.autoBroadcastJoinThreshold |
| spark.locality.wait | spark.sql.inMemoryColumnarStorage.batchSize |
| spark.maxRemoteBlockSizeFetchToMem | spark.sql.inMemoryColumnarStorage.compressed |
| spark.memory.fraction | spark.sql.crossJoin.enabled |
| spark.driver.memory | spark.network.timeout |
| spark.driver.cores | spark.speculation |
| spark.io.compression.codec | spark.serializer |
| spark.io.compression.zstd.blockSize |  |

表2.67方案三tpcds-21G参数列表

|  |  |
| --- | --- |
| **Spark参数** | **范围** |
| spark.executor.cores | 1-3 |
| spark.executor.instances | 4-8 |
| spark.executor.memory | 3-7G |
| spark.executor.memoryOverhead | 384-877m |
| spark.broadcast.blockSize | 32-64m |
| spark.broadcast.checksum | FALSE TRUE |
| spark.broadcast.compress | FALSE TRUE |
| spark.default.parallelism | 200-500 |
| spark.kryoserializer.buffer | 32-128k |
| spark.kryoserializer.buffer.max | 32-128m |
| spark.locality.wait | 4-10s |
| spark.maxRemoteBlockSizeFetchToMem | 1073741567-2147483136 |
| spark.memory.fraction | 0.5-0.9 |
| spark.memory.offHeap.enabled | FALSE TRUE |
| spark.memory.offHeap.size | 512-1024m |
| spark.memory.storageFraction | 0.5-0.9 |
| spark.rdd.compress | FALSE TRUE |
| spark.reducer.maxBlocksInFlightPerAddress | 1073741823-2147483647 |
| spark.reducer.maxReqsInFlight | 2147483646-2147483647 |
| spark.reducer.maxSizeInFlight | 40-72m |
| spark.scheduler.revive.interval | 500ms-100ms |
| spark.shuffle.file.buffer | 16-48k |
| spark.shuffle.io.numConnectionsPerPeer | 1-3 |
| spark.shuffle.sort.bypassMergeThreshold | 150-350 |
| spark.storage.memoryMapThreshold | 1-4m |
| spark.sql.shuffle.partitions | 200-400 |
| spark.sql.autoBroadcastJoinThreshold | 5242880-15728640 |
| spark.sql.inMemoryColumnarStorage.batchSize | 5000-15000 |
| spark.sql.inMemoryColumnarStorage.compressed | FALSE TRUE |
| spark.sql.crossJoin.enabled | TRUE |
| spark.driver.memory | 4g |
| spark.driver.cores | 1 |
| spark.io.compression.codec | zstd |
| spark.io.compression.zstd.blockSize | 32k |
| spark.network.timeout | 120s |
| spark.speculation | FALSE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |

表2.68方案三tpcds-21G参数范围

优化结果：使用的22个敏感query如图2.69所示，使用全部参数随机生成配置文件使测试程序（敏感query）运行成功115次，使用gbdt， ada，xgb，rf建模用参数预测执行时间，筛选重要参数，最终建模误差分别为0.055, 0.066, 0.057, 0.071，建模搜索结果如表2.70和图2.70所示。各模型筛选出来的重要参数见表2.71。将4个模型放入遗传算法，设置种群数量为400，迭代轮次为100，遗传算法结果如表2.72所示，最优配置执行时间为819s，对于tpcds-21G（敏感query）基准配置，执行时间1736s，优化2.1倍，见表2.73，使用遗传算法得出的配置运行测试程序（除query77和query72以外的全部query），运行结果如表2.74和图2.74所示，最优执行时间为2741s，对于tpcds-21G（除query77和query72以外的全部query）基准配置，执行时间5128s，优化1.9倍，优化耗时3813min，见表2.75。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| query14a.sql | query69.sql | query54.sql | query33.sql | query40.sql | query21.sql | query99.sql | query64.sql |
| query20.sql | query95.sql | query62.sql | query43.sql | query50.sql | query29.sql | query15.sql | query23a.sql |
| query23b.sql | query14b.sql | query70.sql | query57.sql | query58.sql | query41.sql |  |  |

表2.69方案三tpcds-21G敏感query

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | gdbt | ada | xgb | rf |
| 模型误差 | 0.055 | 0.066 | 0.057 | 0.071 |
| 重要参数个数 | 9 | 3 | 9 | 4 |
| 遗传算法得出  最优时间 （s） | 729.39 | 897 | 742.03 | 1022.57 |
| 配置一实际  执行时间（s） | 859 | 1881 | 839 | 849 |
| 配置二实际  执行时间（s） | 1191 | 1886 | 834 | 854 |
| 配置三实际  执行时间（s） | 945 | 1801 | 824 | 819 |

表2.70方案三tpcds-21G（敏感query）建模搜索结果

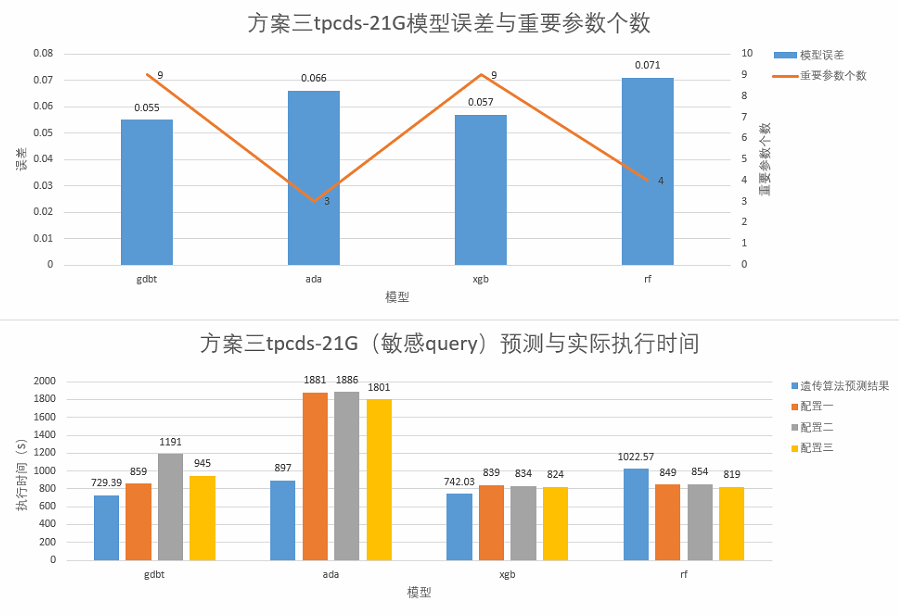


图2.70方案三tpcds-21G（敏感query）建模搜索结果

|  |  |
| --- | --- |
| 模型 | 参数 |
| gdbt | spark.sql.shuffle.partitions |
| spark.executor.memory |
| spark.shuffle.sort.bypassMergeThreshold |
| spark.executor.cores |
| spark.reducer.maxSizeInFlight |
| spark.executor.instances |
| spark.memory.storageFraction |
| spark.shuffle.file.buffer |
| spark.kryoserializer.buffer |
| ada | spark.shuffle.sort.bypassMergeThreshold |
| spark.sql.shuffle.partitions |
| spark.executor.memory |
| xgb | spark.sql.shuffle.partitions |
| spark.executor.cores |
| spark.shuffle.sort.bypassMergeThreshold |
| spark.executor.memory |
| spark.memory.storageFraction |
| spark.shuffle.file.buffer |
| spark.default.parallelism |
| spark.executor.instances |
| spark.storage.memoryMapThreshold |
| rf | spark.shuffle.sort.bypassMergeThreshold |
| spark.executor.cores |
| spark.executor.memory |
| spark.sql.shuffle.partitions |

表2.71方案三tpcds-21G（敏感query）重要参数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | 参数 | 取值 | 执行时间(S) | 预测时间(S) |
| 固定配置 | | spark.driver.memory | 4g | - | - |
| spark.driver.cores | 1 |
| spark.io.compression.codec | zstd |
| spark.io.compression.zstd.blockSize | 32k |
| spark.network.timeout | 120s |
| spark.speculation | FALSE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.sql.crossJoin.enabled | TRUE |
| gdbt | 配置一 | spark.sql.shuffle.partitions | 322 | 859 | 729.39 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 259 |
| spark.executor.cores | 2 |
| spark.reducer.maxSizeInFlight | 67m |
| spark.executor.instances | 8 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 20k |
| spark.kryoserializer.buffer | 128k |
| 配置二 | spark.sql.shuffle.partitions | 321 | 1191 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 264 |
| spark.executor.cores | 1 |
| spark.reducer.maxSizeInFlight | 68m |
| spark.executor.instances | 8 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 21k |
| spark.kryoserializer.buffer | 128k |
| 配置三 | spark.sql.shuffle.partitions | 322 | 945 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 196 |
| spark.executor.cores | 2 |
| spark.reducer.maxSizeInFlight | 67m |
| spark.executor.instances | 7 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 21k |
| spark.kryoserializer.buffer | 128k |
| ada | 配置一 | spark.shuffle.sort.bypassMergeThreshold | 205 | 1881 | 897 |
| spark.sql.shuffle.partitions | 315 |
| spark.executor.memory | 7g |
| 配置二 | spark.shuffle.sort.bypassMergeThreshold | 171 | 1886 |
| spark.sql.shuffle.partitions | 307 |
| spark.executor.memory | 7g |
| 配置三 | spark.shuffle.sort.bypassMergeThreshold | 259 | 1801 |
| spark.sql.shuffle.partitions | 313 |
| spark.executor.memory | 7g |
| xgb | 配置一 | spark.sql.shuffle.partitions | 400 | 839 | 742.03 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 323 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.81 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 213 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 2m |
| 配置二 | spark.sql.shuffle.partitions | 400 | 834 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 324 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.81 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 214 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 1m |
| 配置三 | spark.sql.shuffle.partitions | 400 | 824 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 325 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.83 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 202 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 1m |
| rf | 配置一 | spark.shuffle.sort.bypassMergeThreshold | 153 | 849 | 1022.57 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |
| 配置二 | spark.shuffle.sort.bypassMergeThreshold | 153 | 854 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |
| 配置三 | spark.shuffle.sort.bypassMergeThreshold | 150 | 819 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |

表2.72方案三tpcds-21G（敏感query）遗传算法结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| 819s | 全部参数使用默认值 | 1736s | 2.1 | 3053min |

表2.73方案三tpcds-21G（敏感query）优化结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | 参数 | 取值 | 执行时间(S) |
| 固定配置 | | spark.driver.memory | 4g | - |
| spark.driver.cores | 1 |
| spark.io.compression.codec | zstd |
| spark.io.compression.zstd.blockSize | 32k |
| spark.network.timeout | 120s |
| spark.speculation | FALSE |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.sql.crossJoin.enabled | TRUE |
| gdbt | 配置一 | spark.sql.shuffle.partitions | 322 | 2811 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 259 |
| spark.executor.cores | 2 |
| spark.reducer.maxSizeInFlight | 67m |
| spark.executor.instances | 8 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 20k |
| spark.kryoserializer.buffer | 128k |
| 配置二 | spark.sql.shuffle.partitions | 321 | 3633 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 264 |
| spark.executor.cores | 1 |
| spark.reducer.maxSizeInFlight | 68m |
| spark.executor.instances | 8 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 21k |
| spark.kryoserializer.buffer | 128k |
| 配置三 | spark.sql.shuffle.partitions | 322 | 2741 |
| spark.executor.memory | 7g |
| spark.shuffle.sort.bypassMergeThreshold | 196 |
| spark.executor.cores | 2 |
| spark.reducer.maxSizeInFlight | 67m |
| spark.executor.instances | 7 |
| spark.memory.storageFraction | 0.9 |
| spark.shuffle.file.buffer | 21k |
| spark.kryoserializer.buffer | 128k |
| ada | 配置一 | spark.shuffle.sort.bypassMergeThreshold | 205 | 5462 |
| spark.sql.shuffle.partitions | 315 |
| spark.executor.memory | 7g |
| 配置二 | spark.shuffle.sort.bypassMergeThreshold | 171 | 5367  (query88  失败) |
| spark.sql.shuffle.partitions | 307 |
| spark.executor.memory | 7g |
| 配置三 | spark.shuffle.sort.bypassMergeThreshold | 259 | 5131  (query88  失败) |
| spark.sql.shuffle.partitions | 313 |
| spark.executor.memory | 7g |
| xgb | 配置一 | spark.sql.shuffle.partitions | 400 | 2928 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 323 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.81 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 213 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 2m |
| 配置二 | spark.sql.shuffle.partitions | 400 | 2973 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 324 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.81 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 214 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 1m |
| 配置三 | spark.sql.shuffle.partitions | 400 | 2944 |
| spark.executor.cores | 3 |
| spark.shuffle.sort.bypassMergeThreshold | 325 |
| spark.executor.memory | 7g |
| spark.memory.storageFraction | 0.83 |
| spark.shuffle.file.buffer | 48k |
| spark.default.parallelism | 202 |
| spark.executor.instances | 8 |
| spark.storage.memoryMapThreshold | 1m |
| rf | 配置一 | spark.shuffle.sort.bypassMergeThreshold | 153 | 3014 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |
| 配置二 | spark.shuffle.sort.bypassMergeThreshold | 153 | 2968 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |
| 配置三 | spark.shuffle.sort.bypassMergeThreshold | 150 | 3013 |
| spark.executor.cores | 3 |
| spark.executor.memory | 6g |
| spark.sql.shuffle.partitions | 317 |

表2.74方案三tpcds-21G（除query77和query72以外全部query）优化结果

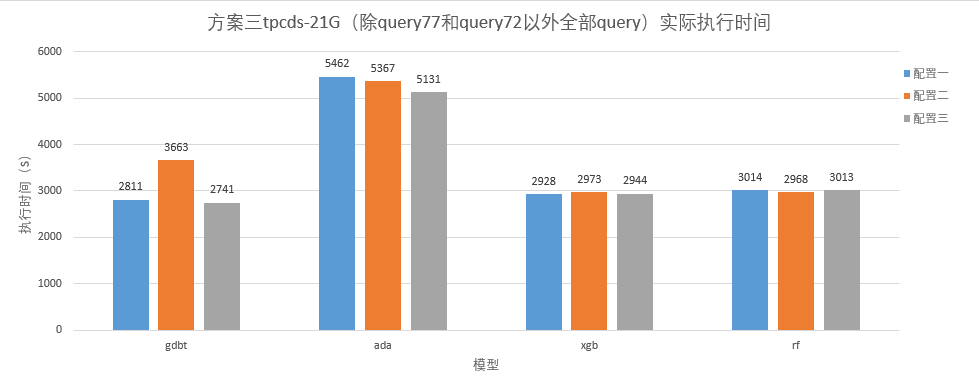


图2.74方案三tpcds-21G（除query77和query72以外全部query）优化结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化总耗时 |
| 2741s | spark.sql.crossJoin.enabled设置为true，其余参数使用默认值 | 5128s | 1.9 | 3813min |

表2.75方案三tpcds-21G（除query77和query72以外全部query）结果汇总

### 2.4.4三种优化方案优化结果对比

优化结果对比如表2.76和图2.76所示，可以看出方案一优化耗时短，在用例上都有较好表现；方案二优化耗时长，优化结果不稳定，在Wordcount-100G上优化结果最好，但在Terasort-20G上优化结果最差；方案三优化耗时短，但在Terasort-20G上优化效果最差。对于Wordcount-100G进行寻找重要参数之后都有了更好的优化结果，但对于Terasort-20G则在寻找重要参数后却得到了更差的优化效果。可能原因是Wordcount-100G有较少的与性能相关的参数，优化方案很容易找出这些重要参数，而Terasort-20G有较多的与性能相关的参数，优化方案很难找出正确的重要参数列表。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | 最优配置执行时间 | 基准配置 | 基准配置执行时间 | 优化倍数 | 优化耗时 |
| Wordcount 100G | 方案一 | 116.824s | 全部参数使用默认值 | 1100s | 9.4 | 1052min |
| 方案二 | 88.895s | 12.4 | 11062min |
| 方案三 | 112.733s | 9.8 | 1258min |
| Terasort 20G | 方案一 | 204.236s | spark.executor.memory设置为5g，其余参数使用默认值 | 1560s | 9.2 | 1934min |
| 方案二 | 169.824s | 7.6 | 22379min |
| 方案三 | 192.389s | 8.1 | 1845min |
| TPC-DS 21G | 方案三 | 2741s | spark.sql.crossJoin.enabled设置为true，其余参数使用默认值 | 5128s | 1.9 | 3813min |

表2.76方案优化结果对比

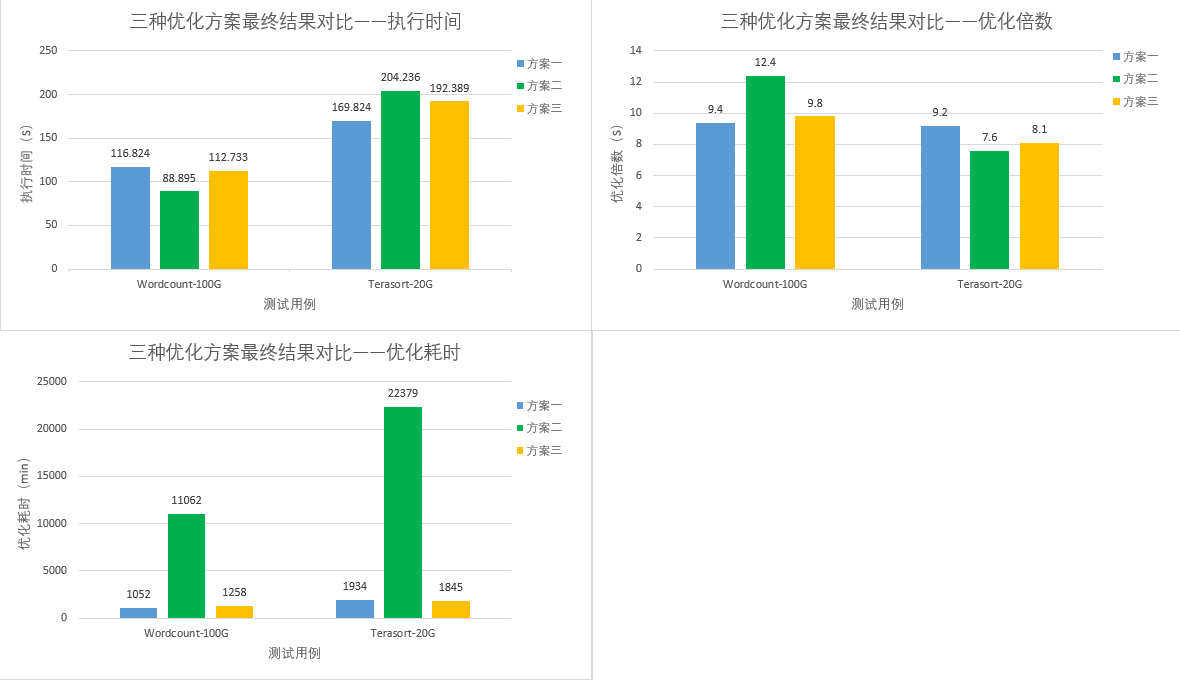


图2.76方案优化结果对比

## 2.5、后续计划

（1）探究自动化参数范围探测方式。

（2）探究敏感query检查方法。

（3）将优化方案使用在Redis上。

（4）探究如何解决大数据量（300G-2T）场景下优化问题。