#### 人员投入

项目投入人员共7人，包括周全（副教授）、关冰宇（博一）、丁志超（研三）、查文睿（研二）、柳端峰（研二）、冯晨旭（研二）、杨雨晴（研二）。

#### 遇到的问题

目前由于设备不足可能造成项目延期。我们将尝试采用GAN网络等方式尽量提高方法执行效率，如期达到验收标准。在技术方面我们可能还需要专家帮助。

#### 未来规划

2021年11月17日 – 2021年11月30日

1. 调优Spark TPCDS基准程序。
2. 尝试找出敏感query的方法。
3. 使用TopDown工具分析Spark基准在容器环境和裸机环境下的特征差异。
4. 调研大数据平台调优论文和相关专利。

2021年12月1日 – 2021年12月15日

1. 采集Redis在容器中的运行特征，根据采集的特征构建预测IPC的性能模型，并给出IPC和执行时间相关的重要性事件以及交互强度。
2. 分析和比较Redis在容器环境与裸机环境下的特征。
3. 调研大数据平台调优论文和相关专利。

2021年12月16日 – 2021年12月31日

1. 通过调研论文和反复实验，探索Redis各个基准程序的合理参数范围。
2. 尝试自动化探索合理配置参数范围的方法。
3. 调优Redis基准程序配置参数。
4. 调研大数据平台调优论文和相关专利。

2022年1月 1日 – 2021年1月15日

1. 采集Kafka在容器中的运行特征，根据采集的特征构建预测IPC的性能模型，并给出IPC和执行时间相关的重要性事件以及交互强度。
2. 分析和比较Kafka在容器环境与裸机环境下的特征。
3. 调研大数据平台调优论文和相关专利。

2021年1月16日 – 2021年1月31日

1. 将容器和裸机环境下的特征分析汇总成一个报告。
2. 调研大数据平台调优论文和相关专利。
3. 根据调研内容尝试方法创新，撰写大数据平台下调优专利。

2022年2月 1日 – 2021年2月15日

1. 大数据集300GB-2TB下Spark/Redis配置调优。
2. 将目前使用过的所有代码汇总起来形成自动化工具。
3. 调研大数据平台调优论文和相关专利。
4. 反复实验并撰写大数据平台下调优专利。

2021年2月16日 – 2021年2月28日

1. 大数据集300GB-2TB下Spark/Redis配置调优。
2. 撰写大数据平台下调优专利。

2022年3月 1日 – 2022年3月15日

1. 将所有调优过的基准程序和优化结果汇总到spark/redis程序性能自动优化技术报告。
2. 撰写大数据平台下调优专利。

2022年3月16日 – 2022年3月31日

1. 输出一套以离线方式自动优化spark/redis程序的软件工具。
2. 输出spark/redis程序性能自动优化技术报告。
3. 输出大数据平台下调优专利。

#### 会议纪要

每周三线上项目例会，每周五实验室内部开会。

|  |  |
| --- | --- |
| 会议日期 | 会议内容 |
| 2021年3月10日 | 调研大数据平台调优、容器特征分析等论文 |
| 2021年3月12日 | 1. 根据容器层、操作系统层、微体系结构层、应用层四个方面调研文章 2. 就目前已读文章，组内讨论进展和文章内容 |
| 2021年3月17日 | 线上讨论已读文章中提及的方法和结果，分析文章参考价值 |
| 2021年3月19日 | 调研spark、redis、kafka性能调优的论文 |
| 2021年3月24日 | 1. 线上讨论已读文章中提及的方法和结果，分析文章参考价值 2. 讨论下一步任务：搭建实验环境，运行指定的基准程序 |
| 2021年3月26日 | 1. 搭建运行环境k8s、spark、docker 2. 调研论文并组内汇报 |
| 2021年3月31日 | 1. 线上讨论坏境搭建问题（系统版本、spark/redis/kafka基准等） 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年4月2日 | 1. 解决环境搭建时的问题 2. 继续搭建环境 3. 调研环境搭建相关论文 4. 调研性能调优论文并组内汇报 |
| 2021年4月7日 | 1. 线上讨论坏境搭建问题（系统版本、spark/redis/kafka基准等） 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年4月9日 | 1. 编写基准运行脚本，根据任务书的测试基准程序 2. 调研环境搭建相关论文 3. 调研性能调优论文并组内汇报 |
| 2021年4月14日 | 1. 线上汇报环境搭建成果和基准运行成功与否 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年4月16日 | 1. 机器持续运行WordCount-50G基准 2. 编写脚本收集程序运行时各个层的事件数据 3. 调研性能调优论文并组内汇报 |
| 2021年4月21日 | 1. 从WordCount-50G开始实验，收集基准运行时的特征 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年4月23日 | 1. 机器持续运行WordCount-50G基准 2. 编写脚本进行数据格式转换、数据清洗、异常值处理 3. 调研性能调优论文并组内汇报 |
| 2021年4月28日 | 1. 尝试构建IPC模型，探索重要事件 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年4月30日 | 1. 构建事件预测IPC的模型，探索重要事件，以所有收集的事件作为输入，输出为IPC，构建集成模型，输出所有事件的重要性排序和比重 2. 调研性能调优论文并组内汇报 |
| 2021年5月5日 | 1. 理清所有采集事件的含义以及该事件是属于哪一层，分析事件和事件之间的关联（交互强度） 2. 开始着手构建交互强度模型 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年5月7日 | 1. 总结所有事件（290多个）的含义 2. 构建事件与事件之间的交互强度模型 3. 调研spark配置参数相关论文 4. 调研性能调优论文并组内汇报 |
| 2021年5月12日 | 1. 开始调研调优spark配置参数的方法，先构建配置参数和重要事件的交互模型来找出重要的配置参数进行调优 2 .线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年5月14日 | 1. 构建GBM计算参数和事件的交互强度模型，分析spark配置参数与重要事件之间的关联 2. 继续调研spark配置参数相关论文，初步筛选配置参数至80多 3. 调研性能调优论文并组内汇报 |
| 2021年5月19日 | 1. 线上讨论构建交互模型时的问题并提出解决方法 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年5月21日 | 1. 对已构建的模型输出结果存在疑问，解决交互强度模型中遇到的问题 2. 按照华为提供的GO语言交互强度模型，对我们的模型进行改进 3. 调研性能调优论文并组内汇报 |
| 2021年5月26日 | 1. 线上讨论交互强度结果 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年5月28日 | 1. 用交互强度模型输出与重要事件相关的重要配置参数 2. 总结我们用到的spark配置参数含义 3. 调研性能调优论文并组内汇报 |
| 2021年6月2日 | 1. 使用spark配置参数预测IPC，找重要参数并与交互强度结果进行比对 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年6月4日 | 1. 构建spark配置参数预测IPC模型，直接找出重要配置参数 2. 将交互强度得出的重要参数与模型直接得出的重要参数比对，总结重要参数的具体含义，分析原因 3. 调研WordCount-100G配置参数的范围 |
| 2021年6月9日 | 1. 尝试调优WordCount100G配置参数。确认WordCount-100G配置参数的合理范围，使用遗传算法搜索基准的最优配置，采用的参数列表为调研论文得出的80个 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年6月11日 | 1. 编写遗传算法搜索配置参数脚本 2. 使用遗传算法脚本对WordCount100G基准使用原参数列表（80多个）进行配置调优 3. 调研性能调优论文并组内汇报 |
| 2021年6月16日 | 1. 尝试先使用遗传算法搜索基准的最优配置，采用的参数列表为我们构建IPC模型预测出来的重要配置 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年6月18日 | 1. 使用遗传算法脚本对WordCount-100G基准使用IPC预测的重要参数进行配置搜索 2. 调研性能调优论文并组内汇报 |
| 2021年6月23日 | 1. 尝试调优WordCount-50G的配置参数，确认WordCount-50G配置参数的合理范围 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年6月25日 | 1. 调研WordCount-50G配置参数的范围  2. 以随机抽样方式运行WordCount50G，收集基准运行时间，将每次运行基准所使用的参数以及最终执行时间记录在表格中，数据格式化保存 3. 调研性能调优论文并组内汇报 |
| 2021年6月30日 | 1. 探索WordCount-50G的重要事件和配置参数 2. 下一步考虑容器配置修改的方法，将其加入spark调优中 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年7月2日 | 1. 构建IPC模型和交互强度模型，探索该基准的重要事件和重要配置参数排序 2. 调研调优容器配置的方法 3. 调研性能调优论文并组内汇报 |
| 2021年7月7日 | 1. 调优WordCount-50G的重要配置参数 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年7月9日 | 1. 根据上周得出的重要参数列表使用遗传算法搜索基准的最优配置 2. 调研调优容器配置的方法 3. 调研性能调优论文并组内汇报 |
| 2021年7月14日 | 1. 探索WordCount-100G的重要事件和配置参数 2. 下一步分析容器环境下和裸机环境下特征的差异性 3. 线上讨论容器配置调优方法 4. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年7月16日 | 1. 构建IPC模型和交互强度模型，探索该基准的重要事件和重要配置参数排序 2. 分别在容器环境下和裸机环境下运行WordCount-50G基准，收集200多个事件的值，1S收集一次，直到基准运行完毕 3. 共一台主节点，三个从节点。对收集到的数据进行数据处理，三个从节点之间取平均值，删除无法采集到的数据，对异常值和零值进行填充或删除 4. 调研性能调优论文并组内汇报 |
| 2021年7月21日 | 1. 调优WordCount-100G的重要配置参数 2. 线上讨论环境差异下特征的波动和稳定 3. 线上讨论调研的容器配置方式 4. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年7月23日 | 1. 构建IPC模型和交互强度模型，探索该基准的重要事件和重要配置参数排序 2. 以绘图的形式分析容器环境和裸机环境下事件差异 3. 调研性能调优论文并组内汇报 |
| 2021年7月28日 | 1. 线上分析和解决不同环境下特征分析方法 2. 线上讨论容器配置修改方式 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年7月30日 | 1. 调研容器中可修改的配置参数及其含义 2. 修改TPCDS运行脚本串行改为并行 3. 重新搭建项目环境，拆分一个服务器出来专门做容器下调优配置的实验 4. 分析容器环境和裸机环境下事件差异 5. 调研性能调优论文并组内汇报 |
| 2021年8月4日 | 1. 线上讨论不同环境下特征的分析结果以及下次改进之处 2. 线上分析和解决当前环境难以修改容器配置问题 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年8月6日 | 1. 继续分析容器环境和裸机环境下事件差异 2. 根据调研结果搭建新的环境用于修改在spark运行时修改容器配置 3. 调研性能调优论文并组内汇报 |
| 2021年8月11日 | 1. 下一步调优terasort20G基准，先找terasort20G的参数范围 2. 线上讨论更改系统环境后仍难以修改容器配置的问题 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年8月13日 | 1. 调研容器的启动过程和容器中参数的修改方式 2. 调研论文和不断实验确定terasort-20G基准的参数范围 |
| 2021年8月18日 | 1. 解决寻找terasort参数范围中遇到的问题，配置参数的合理范围难以确定的问题 2. 线上讨论特征分析结果，改进之处图标的横纵坐标含义描述清楚 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年8月20日 | 1. 调研spark在容器中的提交过程 2. 调研terasort-20G参数确定方法 3. 调研性能调优论文并组内汇报 |
| 2021年8月25日 | 1.构建IPC预测性能模型，探索terasort-20G基准的重要配置参数  2. 尝试修改容器层配置参数，将其加入spark调优 3. 线上汇报本周论文调研结果（思路、创新点、可行度），下次调研论文时重点看A类文章，B以下文章看摘要 |
| 2021年8月27日 | 1. 尝试在k8s上修改容器的配置参数，并运行spark程序 2. 对terasort-20G负载构建性能模型探索重要配置参数 3. 持续收集负载运行时的配置参数和执行时间，需要收集2000个样本才能构建较精确的性能模型 4. 调研性能调优论文并组内汇报 |
| 2021年9月1日 | 1. 讨论目前的调优进度，解决和探讨遇到的问题 2. 将目前已写脚本合成自动化过程 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年9月3日 | 1. 使用随机生成的容器配置，观察对执行时间的改进，得到的结果并不好 2. 调研和解决构建terasort-20G性能模型运行时间过长的问题 3. 继续调研优化论文，侧重点针对运行时间过长的负载怎样快速收敛 |
| 2021年9月8日 | 1. 线上讨论容器配置问题，先尝试使用遗传算法搜索容器的最优配置，看看和默认配置之间的性能提升比例 2. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年9月10日 | 1. 使用遗传算法先搜索容器的较优配置，观察对执行时间的改进 2. 调研性能调优论文并组内汇报 |
| 2021年9月15日 | 1. 尝试使用遗传算法先在80个配置上搜索terasort-20G最优配置参数 2. 线上汇报本周论文调研结果（思路、创新点、可行度） 3. 专利调研 |
| 2021年9月17日 | 1. 为解决构建模型所需样本量过大的问题开展项目交流会 2. 使用遗传算法搜索taresort20G最优配置，观察到使用83个参数导致运行时间过程 3. 根据所有已调研论文进行ABC分类整理，分析论文可行度 4. 针对TPCDS和大数据集下的优化问题，调研优化思路 |
| 2021年9月22日 | 1. 尝试使用遗传算法在交互强度找出的重要参数上搜索terasort-20G最优配置参数 2. 线上讨论容器修改配置后性能无提升的问题 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年9月24日 | 1. 随机跑了60组不同的容器配置，发现这些配置都不如默认配置跑的快，使用遗传算法搜索的最优配置也不如默认配置跑的快 2. 使用交互强度计算出来的36个重要参数，放入遗传算法中进行搜索，最终执行时间的改进效果并不好 3. 找出了在容器坏境下和裸机环境下表现相似的6个事件 |
| 2021年9月29日 | 1. 根据周五讨论会调优terasort-20G配置 2. 继续调研容器更改后的配置比默认配置更差的原因 3. 整理当前的所有事件特征159个，汇总为表（第一列名称，第二列解释）；阅读业界论文对当前表格进行补充 4. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年10月1日 | 1. 根据讨论会的结果更新terasort-20G的配置参数范围，按照一定规则进行分配 2. 构建2个模型：采用小样本迭代的方法构建性能模型，探索重要参数；迭代构建性能预测模型预测给定配置的执行时间，减少遗传算法的搜索时间。 3. 证明容器层参数在默认情况下是不限制资源的，尝试通过截取默认配置下程序多次运行的时间变化以及官方文档描述等 4. 调研发现在没有设置cpuset-cpus的情况下，容器使用了机器的所有cpu,设置以后容器cpu使用被限制，初步验证默认情况下不限制cpu资源 5. 以构建cost模型为思路，找到了另外一种并发运行应用的方法，间接减少运行时间。调研作者的研究方向的历程。 |
| 2021年10月6日 | 1. 尝试调优TPCDS-21G配置参数，首先需找出TPCDS-21G下合理的配置参数范围 2. 调研容器环境下默认参数是不限制资源，配置参数后反而限制资源，达不到优化效果的问题 3. 线上汇报本周论文调研结果（思路、创新点、可行度） |
| 2021年10月8日 | 1. 尝试找出TPCDS-21G运行时的合理参数范围 2. 通过反复在容器上的实验发现--memory-swap在默认情况下是--memory的两倍，测试了从等于memory的值到大于默认值的情况，仍看不到性能的提升 |
| 2021年10月13日 | 1. 线上讨论terasort-20G两阶段模型的构建结果 2. 线上分析容器资源配置问题，下一步从资源配置方面找容器不可调优的依据 3. 线上讨论特征分析结果，下一步调研论文中常常用来分析的指标 4. 对目前为止已读论文按照ABC类、应用环境、方法进行分类 |
| 2021年10月15日 | 1. terasort-20G使用五种继承学习模型构建性能模型，选择精度最高的模型作为最终采用模型并挑出了4个重要配置参数进行调优。使用最终模型和遗传算法结合的方式进行调优，执行时间（112s）相比默认配置（2000s）有了较大提升。 2. 通过实验测量cpu的使用量，在增大允许使用的核数，cpu使用量逐渐提升 |
| 2021年10月20日 | 1. 线上讨论terasort-20G调优结果，开始TPCDS-21G的建模和调优过程 2. 特征分析方面继续调研论文，查找业界常用的特征分析方法和指标 3. 针对容器问题，下一步分析容器资源配置和程序性能之间的影响，通过一次只关注一个指标的变化来分析 4. 根据已读论文提出我们接下来的技术路线，交流并找出方案依据 |
| 2021年10月22日 | 1. 对TPCDS-21G迭代建模探索重要配置参数 2. 找出了一些常用的比较指标。包括cpu(6个), 内存(3个）， 磁盘（10个）， 网络（9个）， 执行时间（1个） 相关的共29个指标。 3. 分析容器资源与性能之间的影响 4. 整理i调研报告和技术路线 |
| 2021年10月27日 | 1. 线上讨论调优过程中遇到的问题，有时一个配置会跑 2. 耗费大量时间。下一步计划：最优点自动化，当模型的误差在阈值次数内趋于平衡，不会再大幅度降低时，停止建模；训练模型的时间不能超过2天；精度误差和TPCDS问题及时交流 3. 继续调研文章，并总结报告技术路线(GAN与贝叶斯、迭代模型)依据 |
| 2021年10月29日 | 1. 修改了迭代建模的脚本，现在如果最近5次迭代的最低误差没有降低的话，脚本就会停止，并且输出目前的误差最低的模型。 2. 找出了3个在所有配置和数据量情况下运行wordCount都不会改变的操作系统层事件  3. 分析得出容器层中cpu设置大于应用层配置时，cpu使用量恒等于应用层配置，运行时间基本相似。但是内存在小于应用层配置时存在一个区间运行成功，其运行时间增大20%，但是内存使用量减少80%。下一步调查容器产生此结果的原因 4. 整理调研报告 |
| 2021年11月3日 | 1. 研究spark参数含义和随机采样策略，定下采样策略，理解之后及时沟通；限制运行时间上限，杀掉的配置直接扔掉 2. 继续调研容器配置，并寻找相关依据 3. 下一步学习并使用TopDown工具进行特征分析 4. 线上汇报调研报告相关内容 |
| 2021年11月5日 | 1. 使用两阶段预测模型和遗传算法优化TPCDS-21G，我们找到的最优配置（2741s）的执行时间相比默认配置（5128s）有了提升 2. 通过看源码研究docker的实现和启用过程 3. 专利调研挑选性能敏感query的问题 4. 完成业界调研报告初稿 5. 编写特征分析报告 |
| 2021年11月10日 | 1. 线上讨论TPCDS-21G调优过程中遇到的问题，query77和query4的spark.sql.shuffle.partitions参数冲突问题；下一步计划：如何自动识别哪个参数导致当前query失败，边跑边确定参数范围 2. 探讨报告书写和格式问题 |
| 2021年11月12日 | 1. 完成特征分析和建模技术报告初稿  2. 完成大数据业务性能优化技术报告初稿 3. 整理目前的所有代码 |