用分批合并与资源监控规划多样本服务器

完整操作教程 · 参考固定 · 往返恢复验收

将多样本合并组织为可核对的批次任务，独立记录最终阶段峰值，避免把分批误当作必然省内存。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

Seurat合并的内存峰值不等于几个RDS文件大小相加。读入时可能同时存在原对象、合并对象和临时副本；稀疏矩阵转换、保留多个assay/layer及并行worker也会改变峰值。规划前先确定研究需要保留哪些数据，再用代表性样本测量，而不是只看文件大小购买内存。

先估算数据和对象规模

按样本记录细胞数、基因数、counts稀疏度、assay/layer、RDS大小和后续步骤。文件大小可帮助规划存储，却不能直接代表R对象内存。尤其不要为估算而对大型稀疏矩阵执行 as.matrix()：如果矩阵变成双精度稠密格式，数值载荷约为行数乘列数再乘8字节，且未包含其他对象和临时空间。

资源项	如何观察	如何用于规划
R对象近似大小	object.size、gc	发现对象是否突然膨胀
进程/作业峰值	作业平台、调度器或系统监控	识别实际运行额度
工作盘占用	df与项目目录阶段性统计	覆盖原始、临时和输出文件
并发影响	记录worker与同机任务	避免把单任务结果当总量

逐步试跑而不是一次复制全队列

先选一对代表对象，在独立目录完成读入、检查、合并和保存；记录每一步的起止时间、退出码、峰值来源、对象大小和工作盘变化。确认counts、metadata、细胞名和关键输出正确后，再增加到预期样本数。每一步只改变样本数或并行数之一，这样才能知道增长来自哪里。

```
date -Is
/usr/bin/time -v -o merge-time.txt Rscript merge_sample.R   > merge.stdout.log 2>
merge.stderr.log
status=$?
printf 'exit_code=%s\n' "$status"
```

GNU Time的最大RSS通常描述被计时进程的观测值，不应直接当作整个作业组或多个worker的聚合峰值。若使用调度器、容器或云平台，优先保存该任务对应的资源统计。保存成功还要重新读入对象，检查维度、assay、layer与metadata。

资源不足时先找重复驻留

若合并阶段内存突然升高，检查原对象是否仍全部驻留、合并副本是否同时存在、是否有不需要的data/scale.data或图，以及并行worker是否复制对象。只有在确认数据结构和保留范围符合研究目标后，才用DietSeurat等方法删减内容；先在副本上核对哪些layer、assay和降维会被移除。counts或模型信息一旦丢弃，未必能从合并对象恢复。

最终资源结论要注明输入规模、版本、保留内容、并发数和测试边界。增加样本、换整合方法或调整layer后应重新试跑。小样本成功只能证明该小样本流程有效；临近正式规模的试跑和结果复读，才可作为租赁规格与期限的依据。

把试跑测量转成容量计划

不要把单个对象的object.size按样本数简单相乘，合并过程的峰值还受临时副本、稀疏度、保留的data层和同时驻留对象影响。记录“读入一个样本”“合并两个样本”“合并后保存/重新读取”三个阶段的峰值和磁盘变化。若样本规模差异明显，试跑应选较大样本或按细胞数分层抽样，并注明抽样方法；小样本只验证流程能执行，无法覆盖超大样本的合并峰值。

若多个同学计划同时运行，先明确同一时间可能重叠的任务数，再在平台额度内做并发试验。单任务成功、两个任务失败，可能表示共享额度或I/O竞争，而不是对象本身错误。将测试日期、节点/容器、R版本、Seurat版本、worker数、退出码和峰值来源写进记录；只有条件相同的数据才适合横向比较。增加内存前，先确认是否能通过按样本顺序读取、及时关闭不再使用的副本或采用磁盘后端满足同一分析目标。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

参考资料

SeuratObject merge.Seurat · 来源资料

Seurat Essential Commands: merge and JoinLayers · 来源资料

Seurat DietSeurat · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。