用对象清单安排分批取数、校验与计算后清理

完整操作教程 · 参考固定 · 往返恢复验收

用版本化对象清单和独立校验值分批下载，计算验收后只清理已确认的工作副本。

运行边界：对象版本、桶和访问凭据须由有权限人员提供；模板不含真实对象或凭据。

对象存储适合保存测序归档，但很多分析程序需要本地文件、随机访问或额外工作区。先按任务取一批数据，下载校验通过再启动分析，通常比把整个桶同步到工作盘更容易控制容量。本篇使用已启用版本管理的Amazon S3通用桶和AWS CLI v2，不假设所有兼容对象存储行为相同。

把批次写成可审查的清单

先生成batch.json数组，每项有name、bucket、key、version、bytes、sha256。name为本批唯一的本地文件名，不包含目录；version是冻结对象版本ID，sha256为源端独立保存的完整文件十六进制校验值。不要把ETag或分片组合校验值直接当完整SHA256。凭据使用已有配置或角色，不写进脚本和清单。

字段	填写规则	检查目的
name	如S1_R1.fastq.gz，单层文件名	避免路径穿越与同名覆盖
bucket/key/version	已确认可读的具体对象版本	防止排队期间读取内容变了
bytes/sha256	实际字节数和64位完整文件校验值	下载完成后独立核对
批次预算	输入、临时文件、结果和余量共存	不能只看压缩FASTQ总量

逐个下载，不覆盖已有目录

将以下代码保存为stage.py，在含batch.json的目录运行python stage.py。清单应先由项目人员核对成对reads、样本对应和对象版本；脚本只解决取数与字节完整性，不验证FASTQ配对及研究范围。下载失败保留部分文件，重试使用新批次目录。

```
import hashlib,json,re,subprocess
from pathlib import Path
rows=json.loads(Path('batch.json').read_text())
assert isinstance(rows,list) and rows
seen=set()
for r in rows:
    name=r['name']
    assert re.fullmatch(r'[A-Za-z0-9][A-Za-z0-9_.-]*',name)
    assert name not in seen;seen.add(name)
    assert r['bucket'] and r['key'] and r['version'] not in ('',None,'null')
    assert type(r['bytes']) is int and r['bytes']>0
    assert re.fullmatch(r'[0-9a-f]{64}',r['sha256'])
root=Path('stage-001');root.mkdir()
for r in rows:
    part=root/(r['name']+'.partial');final=root/r['name']
    subprocess.run(['aws','s3api','get-object','--bucket',r['bucket'],
        '--key',r['key'],'--version-id',r['version'],str(part)],check=True,
        stdout=(root/(r['name']+'.metadata.json')).open('x'))
    with part.open('rb') as f:digest=hashlib.file_digest(f,'sha256').hexdigest()
    assert part.stat().st_size==r['bytes'] and digest==r['sha256']
    part.rename(final)
(root/'DOWNLOAD_VERIFIED').write_text('byte counts and full SHA256 checked\n')
```

把完成标记分成下载和分析两步

DOWNLOAD_VERIFIED只说明整批字节通过检查。启动下游前还需按软件要求检查压缩完整性、FASTQ结构及配对。分析任务另写退出码和产物验收记录；“结果文件出现”不能代替完整任务成功。对象读取受网络、版本权限、加密和归档恢复状态影响，403或InvalidObjectState先查访问和存储类别，不直接归为服务器资源不足。

结果验收后，将必要产物、参数、日志和本批清单归档到确定位置，并从归档位置重新校验。只有归档已确认、源对象版本仍可恢复且保存期允许，才按name清单逐个删除stage-001里的输入工作副本；不对桶执行删除，不用递归通配符清理整个项目。失败任务保留现场直到定位。

例如本批输入80GiB、临时量120GiB、结果20GiB、余量30GiB，示例共需250GiB可用工作空间；这不是实测配比。先用代表任务观察真实峰值，再调整批次，下载中的.partial也计入占用。可以让取数和计算错峰，避免网络与分析I/O争抢同一盘。

验收表按批记录下载状态、样本数、完整字节校验、分析状态、归档位置和清理时间。更换版本或修改清单后重新编号，不沿用旧标记。若流式读取或合理分批仍满足不了必要工作空间，再评估额外工作盘和持续计算环境。

对象版本与清单应能冻结复现

下载前先冻结一次清单，而不是运行中不断扫描并把新对象混进本批。每条对象记录桶名、key、不可变版本ID、预期字节数、独立保存的SHA256、样本ID、数据用途和本地文件名。样本名应可追溯且本批唯一；避免把用户提供的key直接拼接成本地路径。由脚本限制本地文件名必须是安全的单层名称，并校验清单没有重复项、空字段或不符合项目命名规则的项。

对象存储的ETag可能因分段上传或加密方式而不是对象内容的MD5，不能拿它冒充SHA256。AWS S3的VersionId可帮助固定版本，但不同兼容服务的版本与校验能力应以其文档为准。若源端没有可信的完整文件校验值，先在受控的对象导出流程中生成并由数据负责人确认；下载端自行算出的SHA256只能证明当前下载文件是什么，不能证明它与预期源文件相同。

先验证访问和工作盘再启动批次

使用组织提供的角色、配置文件或实例身份，不把访问密钥写到batch.json、命令历史、日志或ZIP。正式批次前检查AWS CLI版本、目标账号与区域、当前凭据身份、桶访问权限、工作盘挂载和剩余空间。使用只读清单时，通过head-object或受控的清单API核对对象版本、长度和元数据；切勿通过列出整个桶来猜测项目范围。传输路径若经过校内代理或跳板机，记录其实际限制。

空间估算不只包括压缩源文件字节数，还要考虑正在下载的临时文件、已经验证的完成文件、下游解压/转换文件、日志和并发批次数。下载到.partial，校验长度及SHA256后才改为最终名。失败留下的partial不要误认为有效输入；下次重试前核对续传行为是否保持对象版本一致。若工具无法保证续传范围属于同一对象版本，删除临时下载后从固定版本重新取更可靠。

分批执行并在计算前设闸门

把manifest切成有限批次，批次大小由工作盘空间、网络带宽、并发和失败重试成本决定。每批先下载，再逐个校验对象版本、长度和SHA256；任何一项不符，就停止该批计算并隔离异常文件。下载和分析同时运行可能让磁盘竞争，也可能令空间峰值增加；先测量串行方案，再决定是否并行。不同样本规模差异大时，可按预计字节或处理阶段分批，而不是只按文件条数平均拆分。

状态表使用明确状态，例如pending、downloaded、verified、analysis_started、analysis_done、failed。写状态时采用临时文件再原子改名，避免中断留下半行记录。分析命令只读取verified清单中的文件，并把样本ID、对象VersionId、校验值和输出目录写入运行日志。批次重跑时识别已完成对象，不能仅凭最终文件存在就跳过验证。

验收下载结果并安全清理

完成后导出本批清单，逐项比较计划对象数、成功下载数、校验成功数、分析输入数和输出数。按样本清单核对配对FASTQ或其他必须成组的文件，防止一端读段缺失。对下游结果运行轻量格式检查和代表样本读取；对象下载成功并不等于文件内容适合分析。日志、manifest和校验清单应保存在受控位置，记录取数时间、工具版本、区域和权限身份摘要，但不记录凭据。

只有确认分析完成、结果已备份、必要审计材料已保存，并且项目负责人同意后，才清理工作盘中的缓存。删除命令必须从清单生成待删路径并先打印核对；不得直接递归删除用户给出的桶key或公共工作目录。对象桶中的源数据和版本保留策略按数据管理要求另行确定，本例只涉及计算工作区副本，不执行云端删除或覆盖。

排查下载失败而不扩大权限

现象	核对顺序	处理原则
AccessDenied	身份、桶/key范围、版本权限和区域	由管理员最小化补充权限，不换个人密钥
长度相符但SHA不符	VersionId、源清单与本地文件	隔离文件，重新取固定版本
下载中断	partial大小、续传版本和网络日志	按工具支持行为续传或重下
工作盘耗尽	并行下载、解压和旧批次共存量	停止新批次，释放经批准的缓存

若工作盘持续不足，可减少批次并发、调整保留策略或比较计算环境中的存储配置，但先确认瓶颈确实是容量，而非权限或网络。服务器租赁也不会自动提供对象桶访问权、数据合规审查或数据分析服务；咨询资源时只需提供任务阶段、文件数量与总字节量、并发、保留周期和所需访问方式，不传递凭据或敏感原始数据。

清单校验器应该先拒绝歧义

取数脚本读取清单后，应在任何网络请求前完成本地验证：必需字段完整、name是单层安全文件名、bucket/key/version均非空、SHA256为64位十六进制、字节数为非负整数、name和样本组合不重复。检测到路径分隔符、`..`、控制字符或重复输出名就终止。清单校验失败时不要开始下载一半再报告错误；在日志中写行号和字段名称，但避免记录访问令牌。

下载器按每个固定VersionId获取对象到唯一partial文件，完成后先核对实际长度，再流式计算本地SHA256，与清单预先提供的值比较。比较成功后在同一文件系统中原子改名为最终文件，并将verified状态写到状态表。若版本ID不受所用兼容服务支持，应将其视为尚未验证的路线，不能静默退回“取当前最新对象”。

断点续传和并发要能追踪

并发数作为明确参数并写入日志；先用一个小批次找到网络、磁盘和服务端的稳定范围，再增加并发。每个下载任务必须有独立临时文件与清晰状态，多个worker不能同时写同一目标文件。中断时保留对象版本、partial大小和工具返回状态。若CLI续传实现不能证明续传部分属于同一对象版本，删除partial并完整重下；不要把不明来源的残缺文件交给下游。

状态表可以记录`sample_id, bucket, key, version_id, expected_bytes, expected_sha256, local_name, status, attempt, started_at, finished_at, error_code`。每次重试新增attempt记录而不是覆盖历史失败原因。凭据身份可记录角色或配置名称与账号末尾标识（经组织允许时），不要记密钥。脚本终止后，人工能据状态表知道哪些文件可用、哪些仍需重试。

计算作业通过独立的verified清单启动，避免通过通配符扫到partial或上次留下的旧文件。下游输出路径也按batch_id分开，运行前检查空目录，运行完成后验证输出数、退出码与代表文件结构。最后将输入清单、校验清单、工具版本和分析日志关联到同一个批次号，方便复现而不暴露凭据。

把对象清单作为分析的输入契约

同一个样本可能在桶内存在多个版本或不同目录下的重复对象。把对象key映射到唯一sample_id和read方向，并在分析前检查R1/R2配对与批次元数据；对象列表本身不证明样本注释正确。按需下载也不意味着可以跳过项目数据管理审批。若本地文件名需要改写，映射表要保留源bucket、key和version_id到本地路径的关系，避免结果文件只显示被截短的文件名而失去来源。

下载后保存一份只读的verified清单，分析脚本从该清单取得输入。不要在分析开始后继续编辑清单；需要补对象时创建新批次号并重算清单哈希。这样同一批分析的输入边界清晰，失败重跑也能判断使用的是哪一版对象。对于生命周期或对象保留策略不明的桶，先与数据管理员确认版本是否会过期，避免分析中途版本被清理。

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

AWS CLI get-object及version-id · 来源资料

S3完整与组合校验值的区别 · 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。