基因容器

POC 测试指导

文档版本	01
发布日期	2019-09-26

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

1 POC 流程说明

GCS POC测试指导手册目标读者为一线技术支持人员，旨在提供step by step的方式指导一线技术支持人员协助基因客户完成基于华为云基因容器服务GCS-CCI方案的POC测试。

指导手册为公司内部文档，未经许可，请不要发送包括客户在内的任何无关人员。

2 业务评估

您需要根据客户业务评估如下几点：

1. 基因分析业务属于第二代测序还是第三代测序？有没有线下使用SGE的经验？

   一般情况下，推荐二代测序业务用GCS-CCI方案或GCS-SGE方案，三代用GCS-SGE方案。本指导主要讲解GCS-CCI方案的测试过程。

   GCS-CCI和GCS-SGE方案的联系和区别：

   - 两个解决方案都是华为云基因容器服务的子解决方案，计算平台底座都是CCI服务，即容器解决方案；
   - GCS-CCI解决方案是华为云自主研发的基因分析解决方案，具有自研语法、流程可视化、资源统计、运维监控、日志、API、SDK等完整的特性生态，但具有少量的学习成本；
   - GCS-SGE解决方案是结合了生信领域传统的SGE集群和CCI容器的优点后做出的基因分析解决方案，具有零学习成本、用户体验和线下一致、支持细粒度资源申请、支持集群节点自动扩缩容等优点，但是目前没有可视化界面和资源监控。
   - 优先推荐有SGE使用经验和有快速上云需求的客户使用GCS-SGE方案；优先推荐具有长远技术和效率追求、对特性生态要求较高的客户使用GCS-CCI方案。

2. POC需要多少资源？

   POC测试时指定区域（Region）需要多少资源，需要与SRE、研发沟通预留资源。此时需要确定此次PoC所在的区域。

3. PoC测试对存储有什么要求？

   根据基因分析的私有工具程序对存储带宽、IO速率的要求，判断使用SFS还是SFS Turbo。对带宽要求高的推荐使用SFS，如生信领域常用的sention软件；对读写性能要求高的推荐使用SFS Turbo。**默认情况下推荐客户使用SFS。**

3 准备工作

3.1 创建 VPC

虚拟私有云（Virtual Private Cloud，以下简称VPC），为云服务器、云容器、云数据库等资源构建隔离的、用户自主配置和管理的虚拟网络环境，提升用户云上资源的安全性，简化用户的网络部署。

操作步骤

步骤1 登录**VPC控制台**，单击“创建虚拟私有云”。

步骤2 根据界面提示配置虚拟私有云和子网参数。

- **区域**：不同区域的资源之间内网不互通。您所有的资源需要在同一个区域创建，请选择**2 业务评估**中确定的区域。
- **网段**：VPC的地址范围，VPC内的子网地址必须在VPC的地址范围内。
- **子网网段**：子网的地址范围，需要在VPC的地址范围内。

  说明

  VPC和子网的掩码设置将直接关系到最大可分配IP的数量，IP数量与基因分析时的并行容器数量正相关，建议IP数分配多一点，**建议掩码配置为16**。

----结束

3.2 获取访问密钥

访问密钥用于访问OBS存储桶时鉴权。

操作步骤

步骤1 登录管理控制台，在控制台页面中单击右上角的用户名，选择“我的凭证”。

步骤2 选择“访问密钥”，单击“新增访问密钥”。

步骤3 输入登录密码，单击“确定”。

访问密钥会自动下载，请妥善保存。

----结束

3.3 创建 GCS 环境

步骤1 登录基因容器控制台，单击“环境管理 > 创建环境”。

步骤2 根据界面填写参数。

- **环境类型**

  选择“云容器示例CCI”。

- **默认环境**

  选择“是”。

- **命名空间选择**

  命名空间是对于同一用户下的云容器实例的逻辑划分，适用于用户中存在多个团队或项目的场景。

  可以选择“已有命名空间”或“新建命名空间”。建议为此环境选择“新建命名空间”，保证资源的隔离独立。

  说明

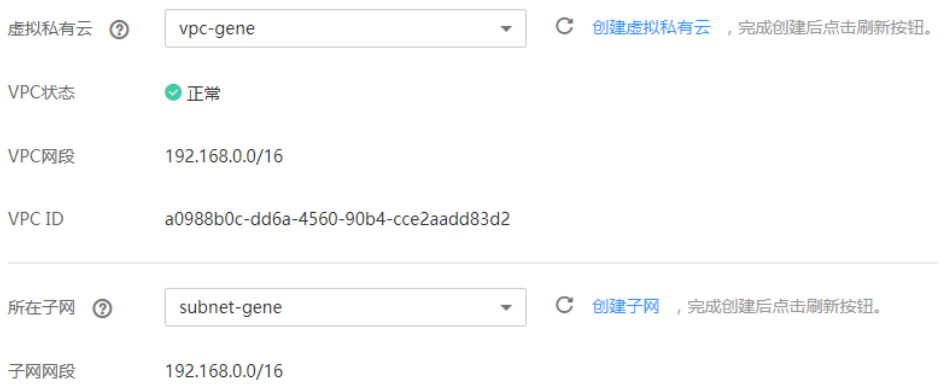
  命名空间名称在云容器实例中需全局唯一，此处请使用自定义名称，以防止被占用。

- **访问密钥**

  上传3.2 获取访问密钥创建的访问密钥。

- **虚拟私有云**

  选择3.1 创建VPC创建的VPC和子网。



步骤3 单击“下一步”，然后单击“提交”。

----结束

3.4 创建并导入 SFS

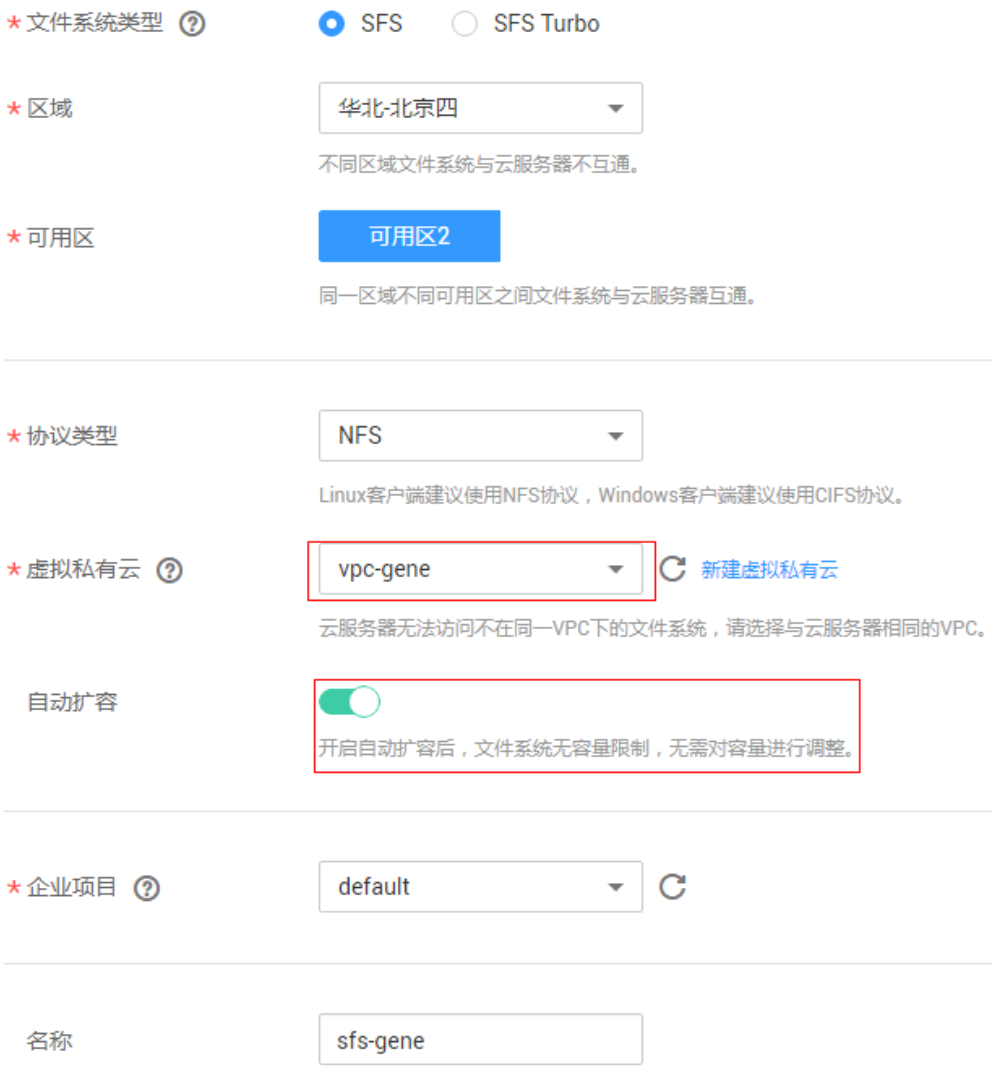
弹性文件服务SFS用于存储基因分析执行过程中产生的中间数据和结果数据。

创建 SFS

步骤1 登录**弹性文件服务**，单击“创建文件系统”。

步骤2 根据界面提示配置参数，**请务必保持打开“自动扩容”按钮**。

- **区域**：租户所在的区域。
- **协议类型**：选择NFS。
- **虚拟私有云**：选择**3.1 创建VPC**创建的VPC。



步骤3 单击“立即创建”。

----结束

在 CCI 中导入 SFS

SFS创建完成后，您需要在CCI中将SFS导入。

步骤1 登录**云容器实例控制台**，在左侧选择“存储管理 > 文件存储卷”，选择**3.3 创建GCS环境**步骤中创建的命名空间。

步骤2 单击右侧“导入”。

步骤3 选择**创建SFS**步骤中创建的SFS，然后单击“导入”。

导入后在“存储管理 > 文件存储卷”页面能够查询到导入的SFS卷。

----结束

4 建立分析流程

以下以GATK流程为例，介绍如何在GCS上创建分析流程。

4.1 流程概述

分析流程是指基因分析需要执行的一系列有序的业务操作的集合。

客户采用传统流程进行样本分析时，一般具有以下特点：

- python或shell脚本编写流程
- 在虚机或物理机上运行
- 业务各步骤不解耦

华为云基因容器服务提供流程设计器进行生信流程设计，具有以下特点：

- 必须使用GCS语法编写，GCS语法为定制化的yaml语法
- 流程由输入、输出、卷和业务四部分组成
- 与客户原有流程兼容，做简单适配即可

4.2 制作镜像并上传

GCS-CCI基因分析方案中，所有的生信工具都是在容器中执行，您需要将基因分析流程中生信工具制作成docker镜像。docker镜像相关知识请参见：**Docker镜像使用|菜鸟教程** 和 **Docker基础和常见操作**。第6章附录中给出了更为消息的镜像制作说明。

镜像制作

制作docker镜像有以下几种方法：

- 通常情况下，成熟、通用的生信软件和工具在Docker Hub上都有现成的镜像，您可以先从**Docker Hub**上搜索查找，请参见：**docker search的使用方法**。

- 若在DockerHub上无法找到满足需求的镜像且无能力自己制作镜像，可以求助华为云容器镜像服务（SWR服务）团队，接口人：李华 l00283074（研发）/徐伟斌 x00417977(SRE)
- 若客户具有一定的IT能力且希望自制镜像，请参见：**Docker镜像的制作方法**。

镜像上传

镜像制作完成后，需要上传至华为云私有镜像仓库。

步骤1 登录**容器镜像服务控制台**。

步骤2 在左侧菜单栏选择“我的镜像”，单击右侧“客户端上传”。

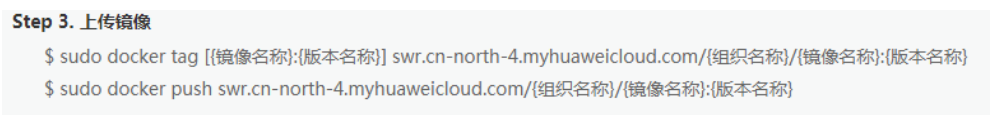
步骤3 在弹出的页面中单击“生成临时docker login指令”，单击 复制docker login指令。docker login指令末尾的域名即为当前镜像仓库地址，记录该地址。

步骤4 在安装Docker的机器中执行上一步复制的docker login指令。

登录成功会显示“login succeeded”。

步骤5 上传镜像。

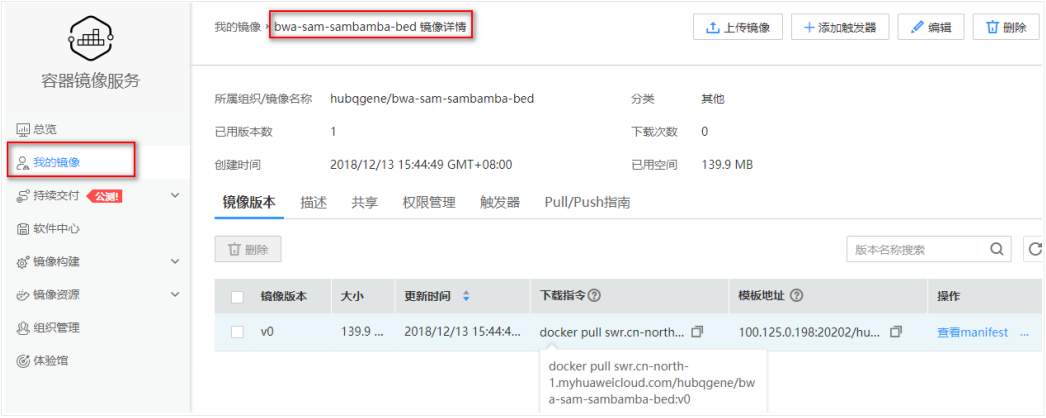
根据界面提示，先给要上传的镜像打tag，然后使用docker push命令上传镜像。



- swr.cn-north-4.myhuaweicloud.com为镜像仓库地址，不同区域有不同的地址，上传镜像时请从界面上拷贝。

- 组织名称可以自定义，首次上传时容器镜像服务会根据组织名称自动创建一个组织。

步骤6 镜像上传成功后，可进入“我的镜像 > 自有镜像”中查看镜像详情。



----结束

添加私有工具

镜像上传到容器镜像后，您需要将上传的镜像添加到GCS工具仓库，以便在基因处理流程中使用。

步骤1 登录**GCS控制台**，选择左侧导航栏的“工具仓库”，在右侧页面单击“添加工具”。

步骤2 设置工具参数，其中带“*”的参数为必填参数。

- * 工具镜像：单击“选择镜像”，选择“我的镜像”，选择**镜像上传**步骤中上传的镜像。
- * 工具名称：新建工具的名称。
- * 工具版本：新建工具的版本。同一工具支持多版本，添加成功后可查看版本情况，如**图4-1**。

图 4-1 工具版本

- 发布者：工具发布者名称。
- 工具LOGO：工具LOGO为60*60px大小，上传图片支持PNG、JPG格式。
- * 标签：工具的类别，可选择已有标签（软件环境、预处理、基因组分析基础、GATK流程、深度学习突变检测、其他），也可添加自定义标签。添加成功后，可在“工具仓库>私有工具”的对应标签下找到该工具。
- 使用说明：工具的说明。
- 常用命令：多条命令以换行区分。

步骤3 单击“完成”，工具即添加成功。可在“私有工具”中查看添加的工具。

----结束

4.3 上传分析数据

OBS桶用于存储基因分析流程要分析的原始数据，原始数据由客户提供。

创建 OBS 桶

步骤1 登录**OBS控制台**，单击右侧“创建桶”，在弹出页面中填写相关参数，如下图所示。

- 区域：选择POC所在的区域。
- 桶名称：

步骤2 单击“立即创建”。

----结束

上传数据

步骤1 登录GCS控制台，在左侧选择“数据存储 > 私有数据”。

步骤2 请在“数据存储”中为原始数据建立一个文件夹。

步骤3 单击工具上传，根据界面提示下载并安装OBS客户端，上传数据。

上传完成后，在OBS客户端中查看如下结果，表示上传成功，其中“对象”列表示数据在OBS桶中的存放路径。

数据上传成功后，您可以进入“数据存储 > 私有数据”，选择上传的OBS桶，进入对应目录下查看已上传的测试数据。您还可以对测试数据进行下载和删除操作。

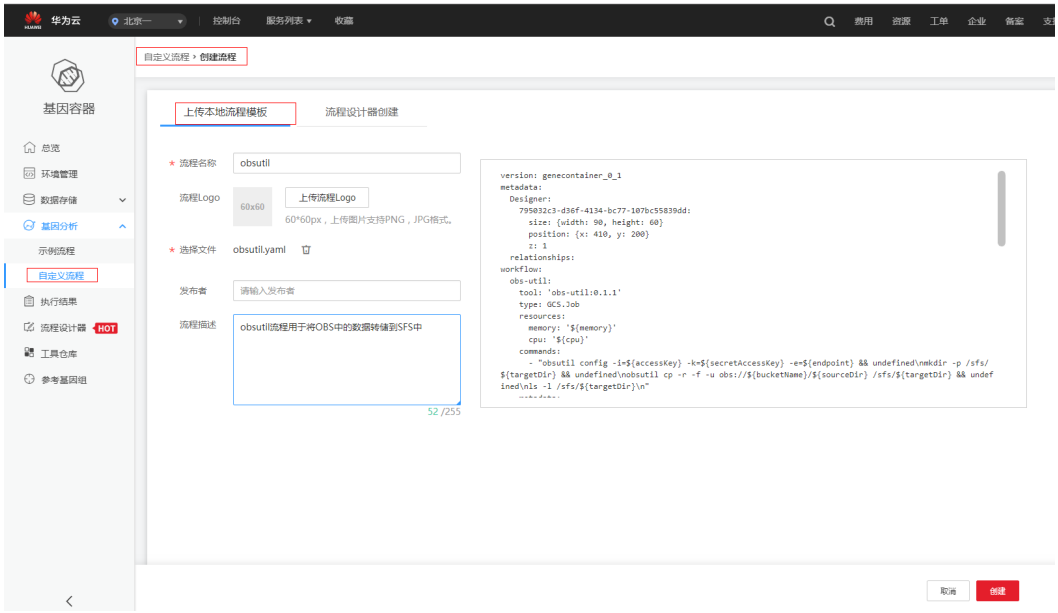
----结束

转储数据

步骤1 和基因容器团队联系，索要obsutil流程（obsutil.yaml文件），该流程用于将客户OBS中的数据转储到SFS中，接口人：吴雷 w00445106。

步骤2 将obsutil.yaml文件通过本地上传的方式导入为自定义流程。

图 4-2 上传 obsutil 流程



步骤3 使用obsutil流程将OBS桶中指定目录下的样本源数据和reference数据转储到SFS的指定目录下，提交任务，等待任务执行完成。

图 4-3 自定义流程列表中找到 obsutil 流程，开始分析

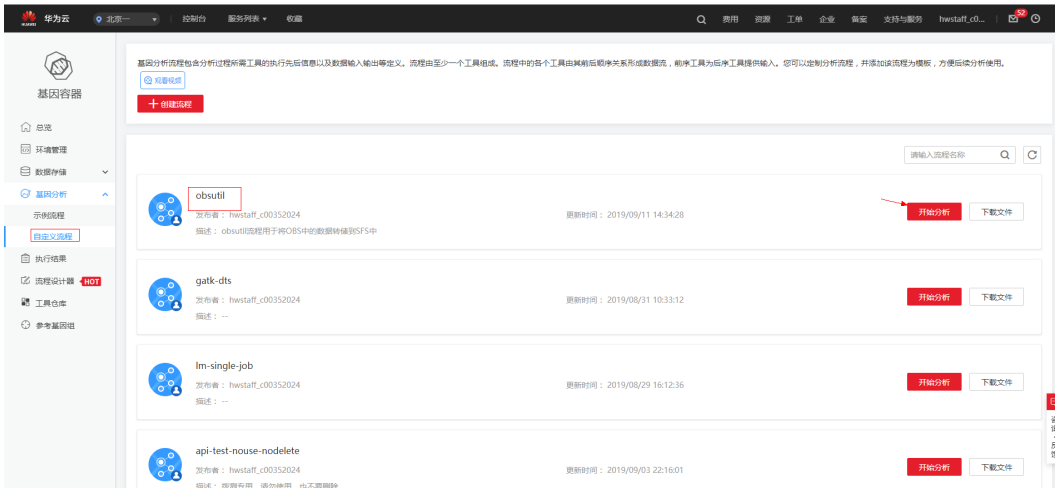


图 4-4 配置本次转储任务的基本信息和目标 SFS

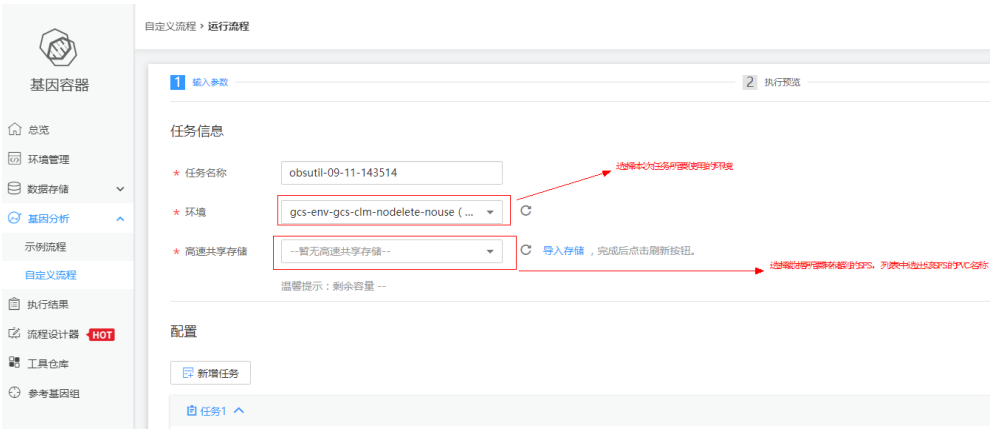
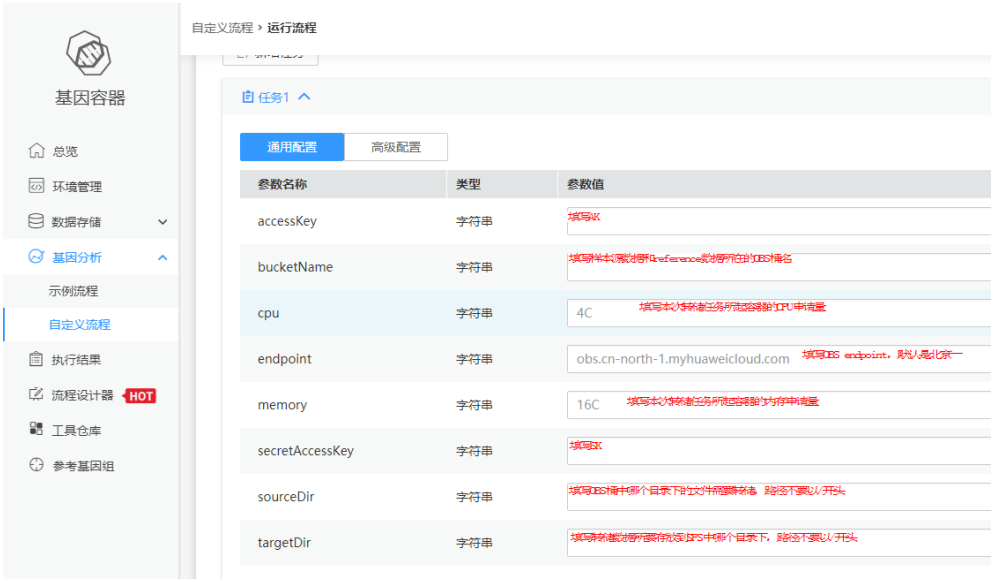


图 4-5 配置转储任务参数



进行以上配置时请注意以下事项：

- “高速共享存储”中选择的是目标SFS的PVC名称，即数据转储到哪个SFS
- 参数列表中AK/SK配置桶拥有者账户的AK/SK，一般为当前账户的AK/SK
- 配置OBS桶源路径sourceDir和SFS中目标路径targetDir时，均为相对路径，即不要以“/”开头
- 该流程仅支持OBS桶中指定路径下的文件批量转储到SFS指定路径，不支持单个文件转储，也不支持指定路径下的部分文件的批量转储。
- 配置memory参数时，请保证给定的值大于所要转储的最大单个文件大小，防止偶现文件转储不完整的情况。
- 当前obsutil流程持续更新，后续所发布新版本将及时推送大家。

----结束

4.4 流程设计

生信流程设计共包括6步：

1. 业务梳理
2. 框架搭建
3. 输入变量抽取
4. 卷设置
5. 代码编写
6. 流程调试

进行流程设计时，需要使用到基因容器服务提供的流程设计器。

图 4-6 流程设计器入口

4.4.1 业务梳理

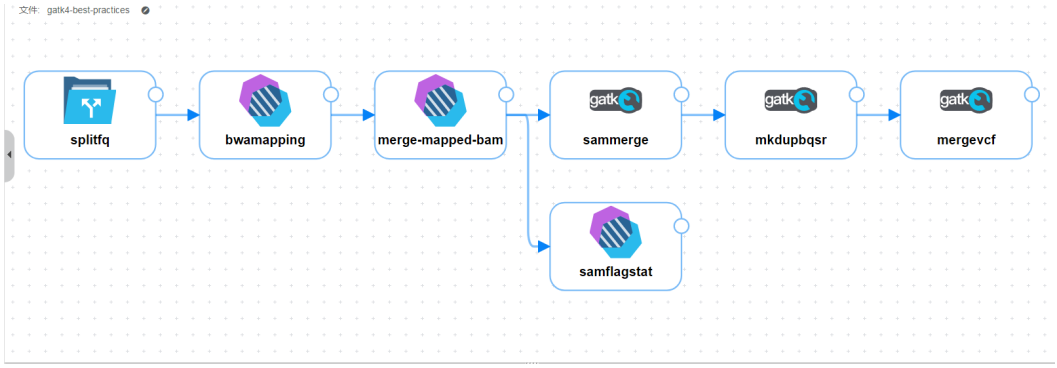
业务梳理是指将客户线下业务流程进行分步解耦，以便后续在GCS上进行设计。梳理时请遵照以下原则：

- 各步骤逻辑解耦，整理出前后依赖关系；
- 梳理各步骤需要用到的生信工具；
- 预估各步骤所需要的CPU和内存资源。

4.4.2 框架搭建

框架搭建是指根据4.4.1 业务梳理中梳理出的解耦步骤，通过拖拽的方式建立流程雏形的过程。以生信领域常用的GATK流程为例，大致分为以下7步。

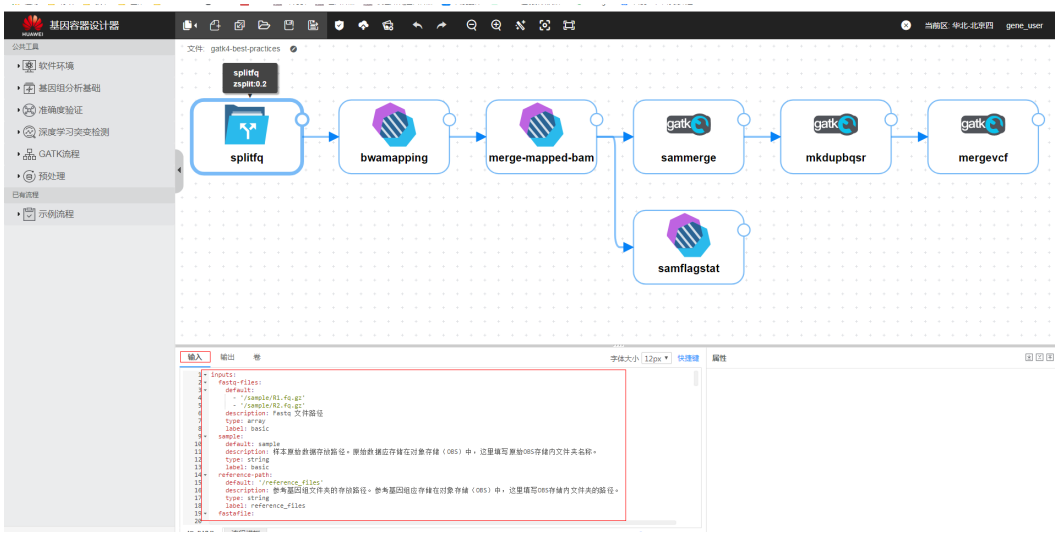
图 4-7 GATK 流程框架



4.4.3 输入变量抽取

变量抽取是指提取出整个生信流程用到的**全局变量**，并定义在输入参数中的过程。在执行某个样本分析前，输入变量的值将由客户赋值，这样流程可以做到参数可调、多次复用。例如，可以将流程中某个步骤计算时申请的CPU和内存抽取成输入变量，这样投递不同的样本分析任务时，客户可以根据当前样本的实际情况定制化给这两个参数赋值。

图 4-8 设计器中输入变量定义区域



字段解释

定义输入变量常用的属性包括type、default、label和description。

- **type**：变量类型，必选字段，默认为string，其他可用类型还有number、bool、array。
- **default**：变量默认值，可选字段。
- **label**：变量的标签，可选字段，默认为“basic”，对客户业务没有任何影响，用于逻辑分组，增强可读性。
- **description**：变量的描述信息，可选字段，对客户业务没有任何影响，便于设计人员了解变量的用处。

以GATK流程中的reference-path变量为例进行说明。

图 4-9 字段说明样例

名为“reference-path”的变量，是一个string类型的变量，默认值是“/reference_files”，当客户不指定其他值时将使用该值作为变量值。这个变量的label是“basic”，该变量用于记录reference数据在OBS桶中的存放路径。

变量抽取原则

输入变量可用于整个流程，类似于编程语言中的全局变量，用户可根据业务需要决定抽取哪些变量。一般情况下，以下场景多抽取成输入变量：

- 各步骤使用的CPU和内存资源。不同样本时，同一步骤用到的CPU和内存往往不同，可让客户根据经验值进行自定义。不同步骤若使用的资源相同，可统一使用某个变量。
- 文件路径，尤其是reference数据和源数据路径。抽取成变量的文件路径使得流程能更加灵活的适配不同情况下的分析需求。
- 业务代码中多次出现的值或经常变更的值。类似于编程语言中变量的使用场景。

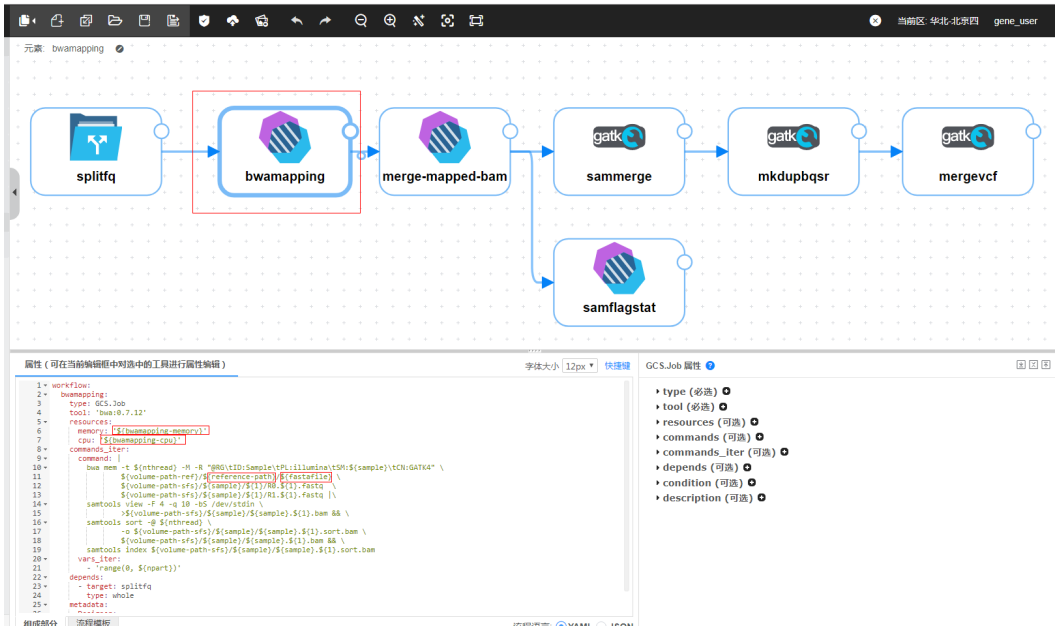
样例

以GATK流程为例，输入变量定义如下：

图 4-10 输入变量定义样例

GATK流程中共定义了fastq-files、sample、reference-path、fastafile、knownsitedb147等多个变量。其中fastq-files为数组变量，其他为字符串变量。

图 4-11 输入变量使用举例



以流程中第2步“bwapping”为例，该步骤的CPU和内存申请的资源量分别由输入变量“bwamapping-cpu”和“bwamapping-memory”的值决定（bwamapping-cpu和bwamapping-memory这两个变量的定义在图3中没截取出来）。变量“reference-path”和“fastafile”被用于业务代码中。使用输入变量时，和shell的用法完全一样，即“${变量名}”。

完成并保存当前流程后，当执行某个样本的分析任务时，在投递任务前可以在“配置”中对所有参数赋值。

图 4-12 投递任务时为输入变量赋值样例

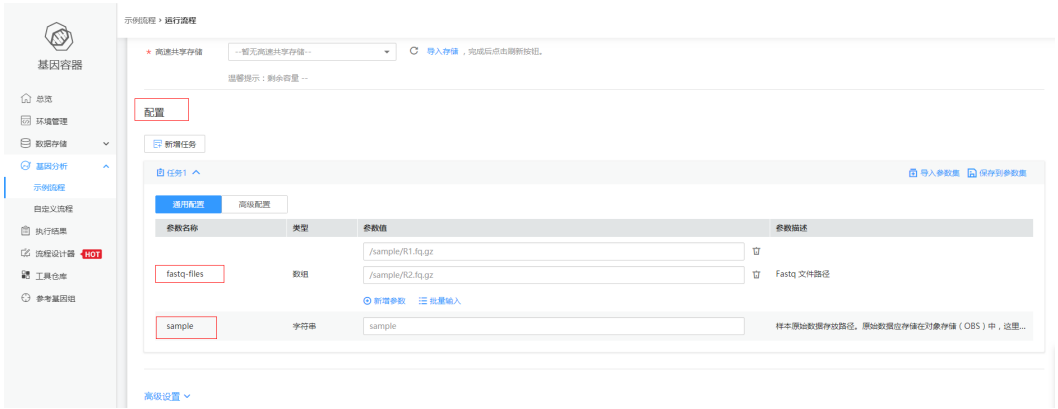
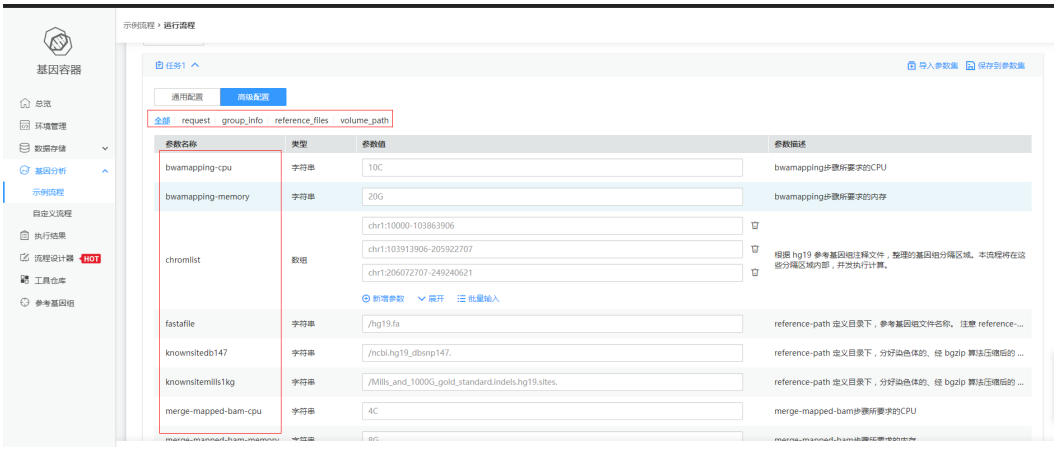


图 4-13 投递任务时为输入变量赋值样例



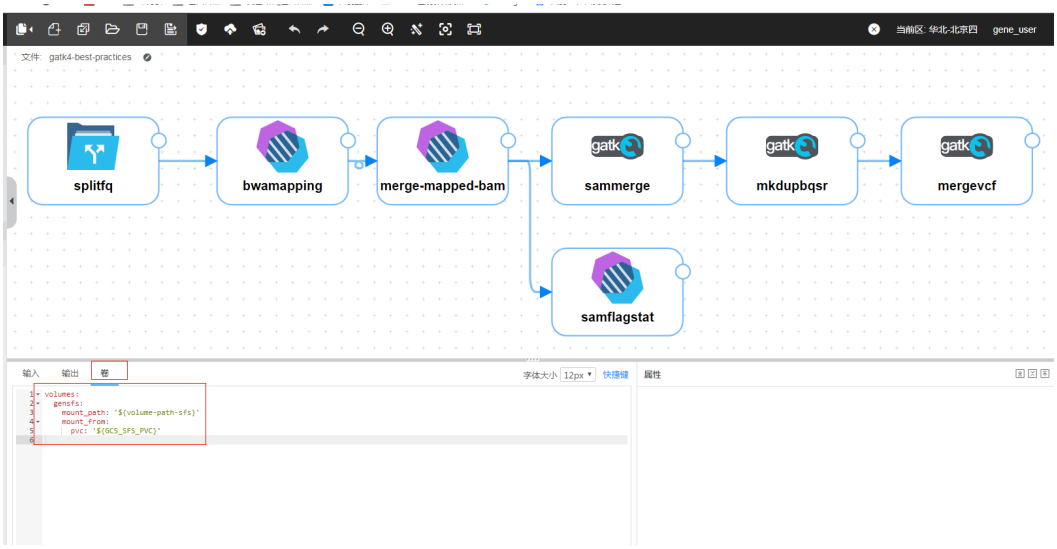
4.4.4 卷设置

在使用基因容器服务时，用户的reference、源数据、中间数据和结果数据都将写入SFS中。如何将SFS挂载到容器的文件系统中，就是卷设置需要解决的问题。容器对卷的挂载操作完全等同于Linux系统的mount挂盘操作。

卷设置中进行了规定了以下设置项：

- 挂载哪块SFS盘；
- 该SFS盘挂载到容器文件系统的哪个路径下。

图 4-14 卷设置



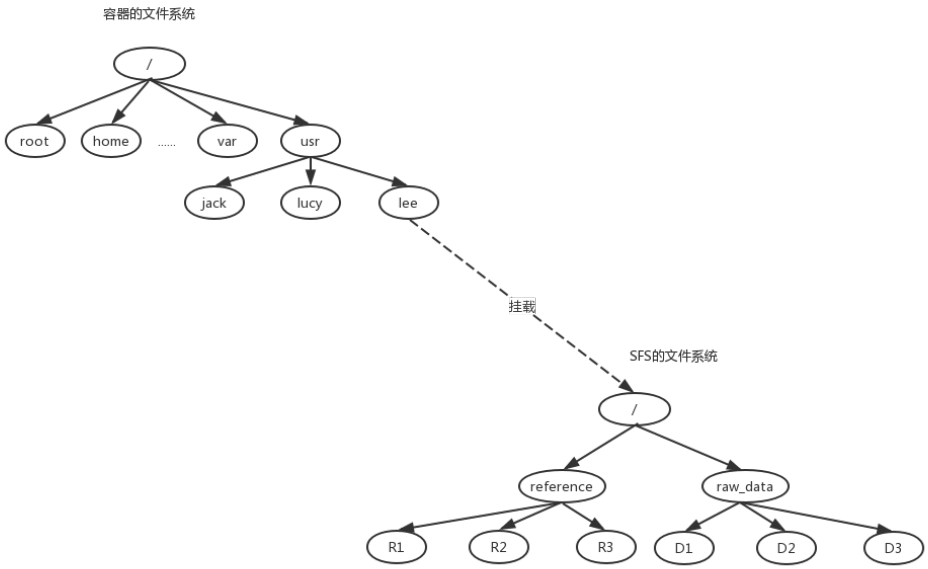
字段解释

卷设置中常用的字段包括mount_path和mount_from。

- **mount_path**：SFS盘挂载在容器文件系统的路径，必选字段，类似于linux系统中mount的挂载点路径。
- **mount_from**：所要挂载的SFS盘的PVC名称，必选字段，该值必须是默认值“GCS_SFS_PVC”，**不可修改！**

挂卷原理

图 4-15 SFS 挂载原理



容器挂载SFS盘有以下要点：

- 容器和SFS中都有一套文件系统目录树；
- 将SFS挂载到容器中时，本质上是将SFS的目录树挂载到容器的目录树中，形成一颗新的目录树；
- 挂载完成后，容器中可直接访问到SFS中的资源，比如在容器中路径/usr/lee/reference/R1即可访问到SFS中R1。

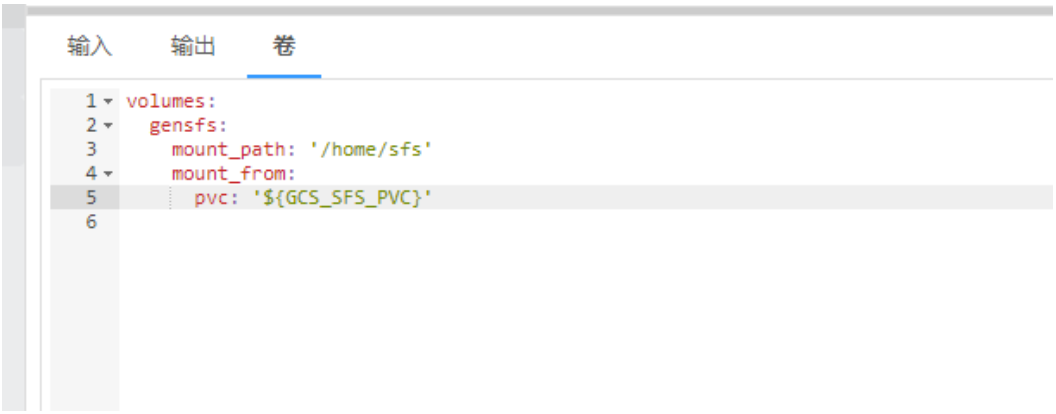
注意事项

- 以“GCS_”开头的变量为GCS服务内置变量，不可修改
- **mountpath的最后一级子目录强烈建议设置为一个新目录，默认为“**/**/.../sfs”，防止挂载卷时因设置为已有目录引起的最后一级子目录下其他子目录被强制覆盖的情况。**

举个例子：若mountpath设置为/var/log，SFS的文件系统将被挂载到该挂载点下，原系统自带的/var/log下其他目录和文件将被清空。若mountpath设置为/home/sfs，由于系统目录/home下原本没有子目录sfs，在挂载卷时将首先在/home下创建子目录sfs，然后将SFS卷挂载至/home/sfs。

样例

图 4-16 卷设置样例



以图4-16设置为例，mount_path指定了将SFS盘挂载到容器文件系统的"/home/sfs"目录下，其中sfs目录将会在挂载的时候被新建；mount_from的子属性GCS_SFS_PVC将决定挂载哪个SFS盘，无需更改，直接复用即可。所以，当在容器中需要访问SFS中的资源时，访问路径前缀为“/home/sfs/”，其后跟的子路径为资源在SFS中的路径。

指导客户进行卷配置操作时，请直接复制下方代码，修改mount_path字段即可，其他不做修改：

volumes:

gensfs:

mount_path: '/sfs/'

mount_from:

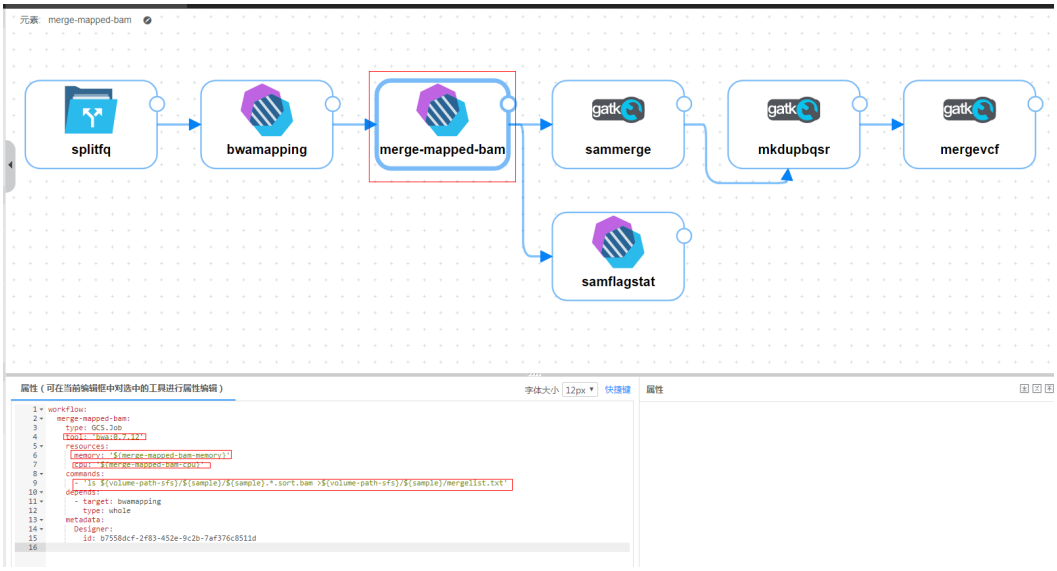
pvc: '${GCS_SFS_PVC}'

4.4.5 代码编写

代码编写是指分步骤的编写4.4.2 框架搭建中规划的业务代码。

以GATK流程的第三步“merge-mapped-bam”为例进行说明。

图 4-17 GATK 流程中第三步“merge-mapped-bam”代码示例



字段解释

- **type**：必须为“GCS.Job”。
- **tool**：该步骤需要用到的私有工具名称和版本号，也是这一步起的容器用到的镜像，格式为“私有工具名:版本号”，可在私有工具列表中获取。
- **resource**：该步骤起容器所需要的资源规格，包括CPU和内存。CPU:Memory要维持在1:2~1:8之间，定义的资源数值必须为0.25的整数倍。
- **commands**：该步骤容器启动后需要执行的业务代码，这一项是最重要的配置。

注意事项

- tool中设置时不要遗漏版本号
- resource中CPU的值不要遗漏单位“c”，memory的值不要遗漏单位“G”，如下示例

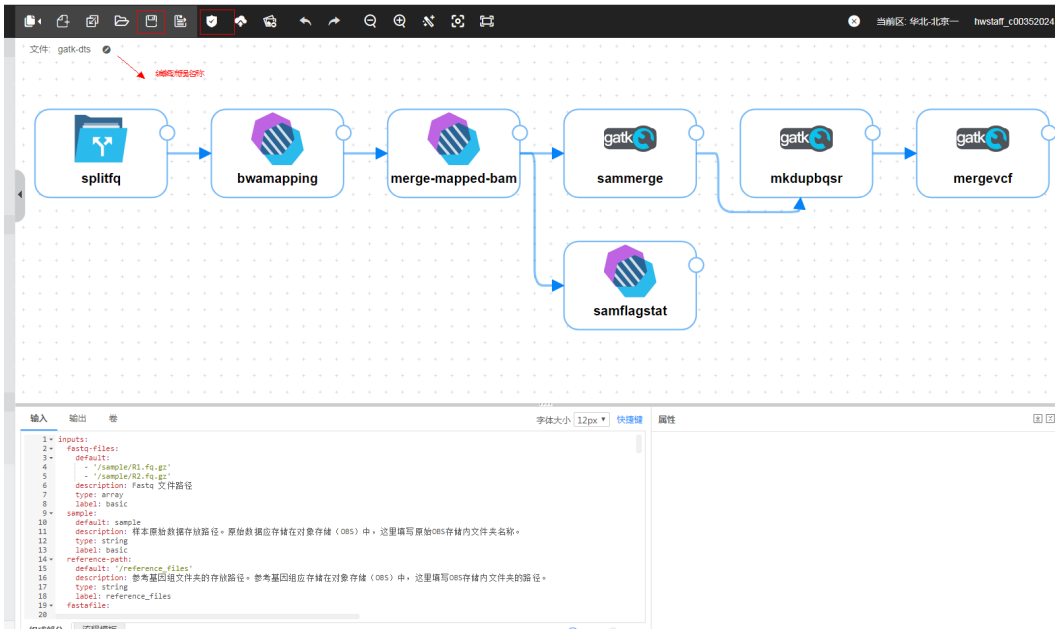
参数名称	类型	参数值
bwamapping-cpu	字符串	10C
bwamapping-memory	字符串	20G

- commands中的写法模仿示例流程即可。对于业务代码较长的，建议将代码放在一个脚本中，脚本置于SFS中某个指定路径下，commands中命令直接调用该脚本。
- 无论流程有多少个步骤，每个步骤均按照以上设置即可。涉及到复杂业务的，如多进程并行、复杂依赖关系等，请联系基因容器团队。

4.4.6 流程调试

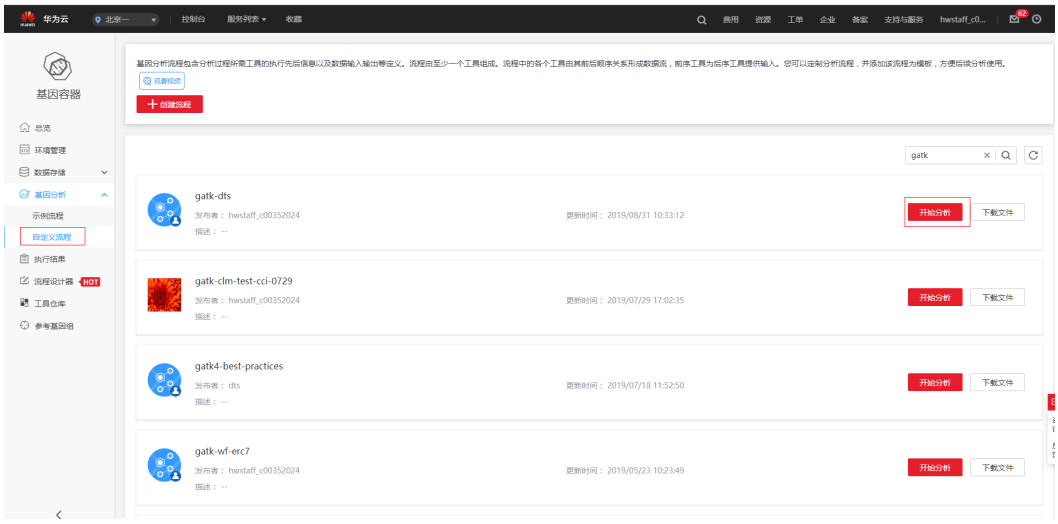
当流程写完后，为流程取个名字，单击“验证流程”按钮可进行流程语法校验，若出现报错可根据报错信息进行修改；单击“保存”，可保存当前流程。

图 4-18 流程的保存和语法校验



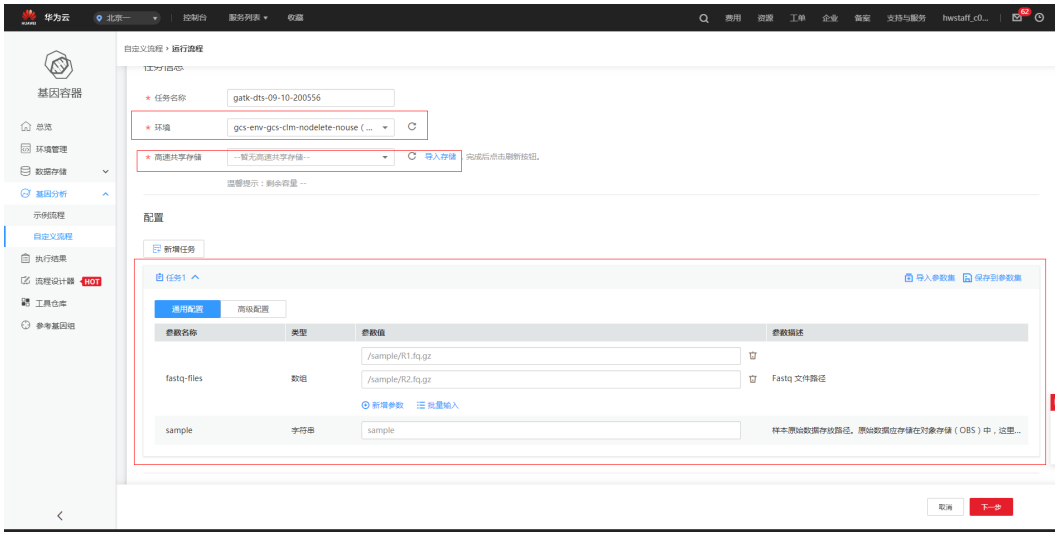
保存完毕后，可通过一个真实的样本验证业务代码的正确性。返回基因容器控制台主页面，单击左侧导航栏中的“基因分析 > 自定义流程”，找到刚才写好的流程，点击右侧“开始分析”。

图 4-19 流程调试 1



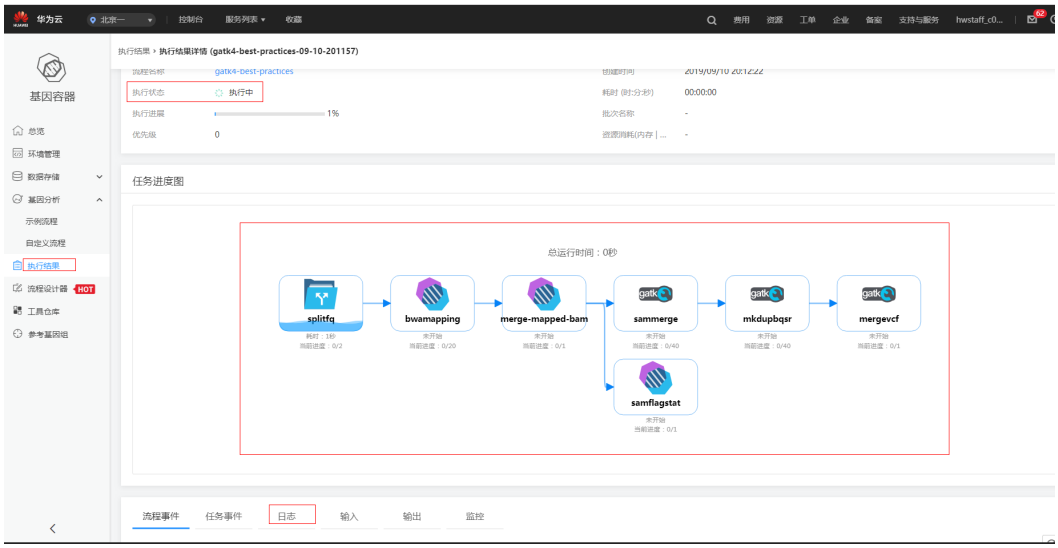
进入分析参数设置页面后，选择用到的分析环境，“高速共享存储”中选择**3.4 创建并导入SFS**步骤中导入的SFS卷，在“配置”中为流程的输入参数赋值，然后单击“下一步”，再单击“提交”即可开始分析。

图 4-20 流程调试 2



样本开始分析后，在执行结果列表中可以看到处于“执行中”的执行结果，点击名称进入详情后，可以通过流程的水纹线观察流程执行进度。若出现标红的步骤，说明流程报错停止，可点击下方的“日志”查看详细报错。

图 4-21 流程调试 3



5 流程测试

5.1 端到端测试

端到端测试是指按照**3 准备工作**和**4 建立分析流程**的内容进行一个样本的完整分析测试，了解基因容器服务的基本使用方法，验证分析的正确性。

5.2 压力测试

压力测试是指在端到端测试完成后，客户使用基因容器服务的PythonSDK、API或CLI进行多个样本分析任务批量分析的测试，考察基因容器服务的并行运算能力和效率。一般情况下，引导客户基于PythonSDK进行压力测试，请直接参考基因容器服务帮助文档：**华为云基因容器服务PythonSDK使用指南**

6 附录

6.1 权限开通

本章节指导您为IAM用户开通GCS权限，如果您是使用账号，请跳过本章节，因为账号具有所有服务的权限。

表 6-1 使用 GCS 服务所需权限

名称	说明
GCS Administrator	基因容器服务管理权限
CCI Admin	云容器实例服务管理权限
OBS Operator	对象存储服务操作权限

说明

CCI Admin和OBS Operator为细粒度权限策略，需要申请细粒度权限公测，具体方法请参见**申请细粒度访问控制公测**。

操作步骤

用户组是用户的集合，IAM通过用户组功能实现用户的授权。您在IAM中创建的用户，需要加入特定用户组后，用户才具备用户组所拥有的权限。关于创建用户组并给用户组授权的方法，可以参考如下操作。

说明

如果您使用已有用户组，可直接从**步骤5**开始授权。

步骤1 使用注册的华为云账号登录华为云，登录时请选择“账号登录”。

图 6-1 登录

步骤2 进入华为云控制台， 控制台页面中单击右上角的用户名，选择“统一身份认证”。

图 6-2 统一身份认证

步骤3 在统一身份认证服务的左侧导航空格中，单击“用户组”>“创建用户组”。

图 6-3 创建用户组

步骤4 在“创建用户组”界面，输入“用户组名称”，以“开发人员组”为例，单击“确定”。

用户组创建完成，界面自动返回用户组列表，列表中显示新建的用户组。

步骤5 单击用户组右侧的“权限配置”。

图 6-4 权限配置

步骤6 在区域所在行，例如“华北-北京四”，单击“设置策略”。

图 6-5 设置策略

步骤7 在“设置策略”中搜索“GCS”，勾选“GCS Administrator”。

图 6-6 设置策略

步骤8 再搜索“CCI”，选择“CCI Admin”，然后单击“确定”。

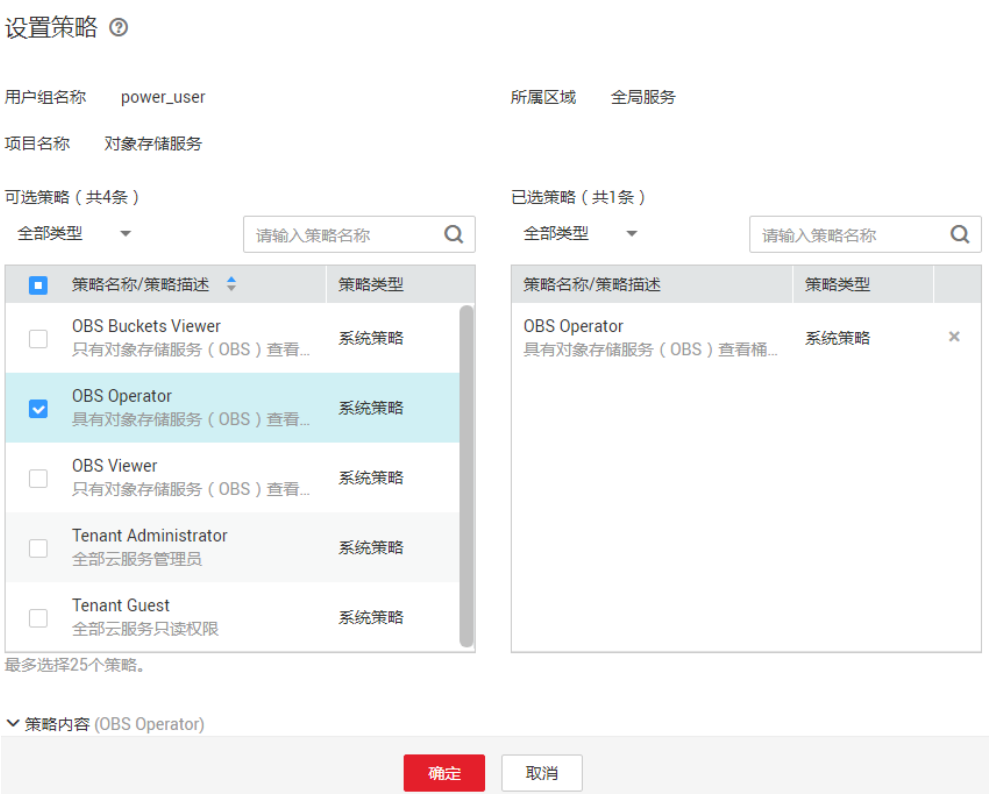
图 6-7 设置策略

步骤9 单击“确定”，完成用户组授权。

步骤10 在在“全局区域 对象存储服务”所在行，单击“设置策略”，如下图所示。

图 6-8 设置策略

步骤11 勾选“OBS Operator”，然后单击“确定”。



----结束

6.2 制作 Docker 镜像

6.2.1 安装 Docker

Docker是一个开源的引擎，可以轻松的为任何应用创建一个轻量级的、可移植的、自给自足的容器。容器镜像服务兼容原生Docker，支持使用Docker CLI和原生API管理容器镜像。

在安装Docker前，请了解Docker的基础知识，具体请参见 **Docker Documentation**。

Docker几乎支持在所有操作系统上安装，用户可以根据需要选择要安装的Docker版本，具体请参见**https://docs.docker.com/engine/installation/**。

说明

- Docker镜像的的存储可以使用华为云提供的容器镜像服务，由于容器镜像服务支持Docker 1.11.2及以上版本上传镜像，建议下载对应版本。
- 安装Docker需要连接互联网，内网服务器需要绑定弹性IP后才能访问。

另外，在Linux操作系统下，可以使用如下命令快速安装Docker。

```
curl -fsSL get.docker.com -o get-docker.sh
sh get-docker.sh
```

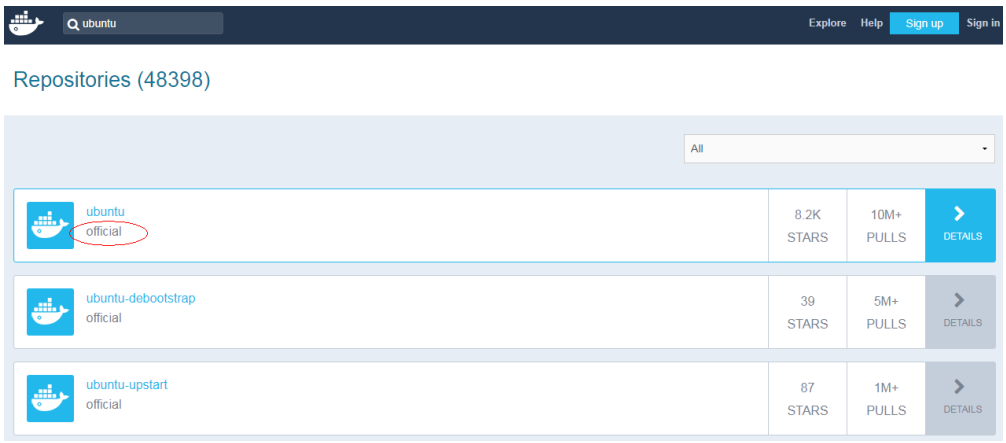
6.2.2 从 DockerHub 搜索目标软件的 Docker 镜像

DockerHub网站提供了40万+的各类软件的公开Docker镜像下载，并以每周5K的速度在持续增长。所以除了是自己开发的软件，一般都可以在这里找到对应的镜像版本。DockerHub地址为**https://hub.docker.com/**。

有以下几类软件，建议您从DockerHub获取Docker镜像，而不是自己制作。

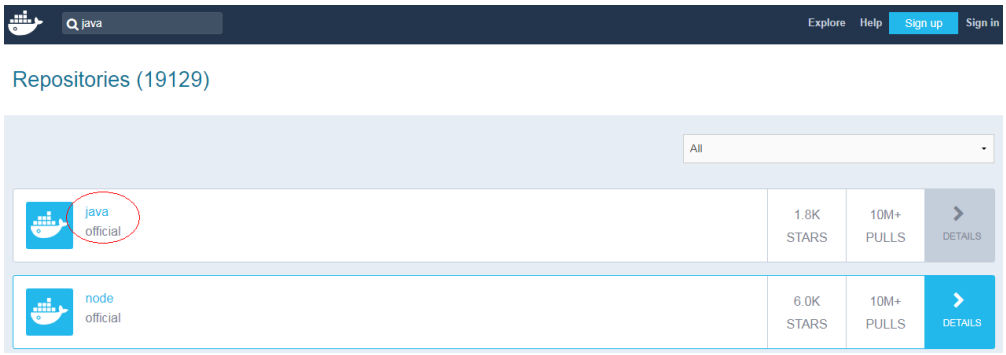
- **基础运行OS环境**

  基础OS类的镜像，例如Ubuntu，Suse，Centos等，建议您直接从DockerHub获取官方认证版。



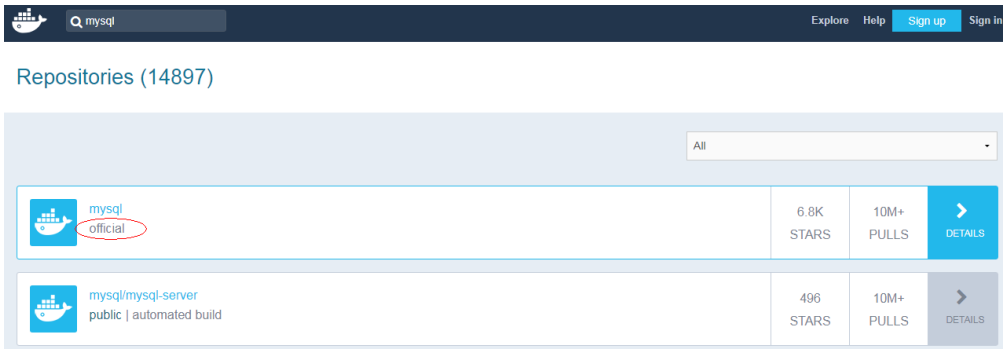
- **基础编程语言类**

  基础编程语言类的镜像，例如Java，Python，R语言，Golang等，建议您直接从DockerHub获取官方认证版。



- **基础通用类软件**

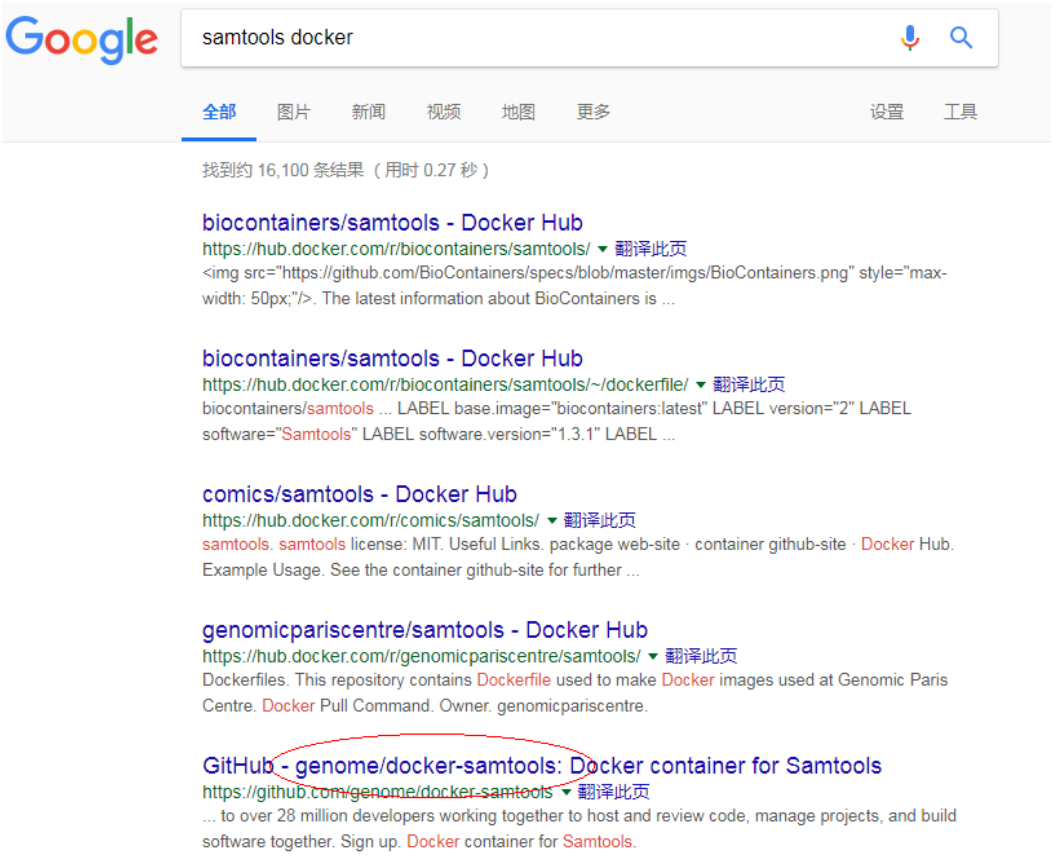
  常见的通用类的软件，例如Tomcat，Mysql，Ngnix等，建议您直接从DockerHub获取官方认证版。



6.2.3 从 Google 搜索目标软件的 Docker 镜像

如果有些软件位于第三方的镜像仓库中，您可以通过Google搜索来查找相关镜像。搜索时只需在软件名后面加上docker关键字即可。

例如：



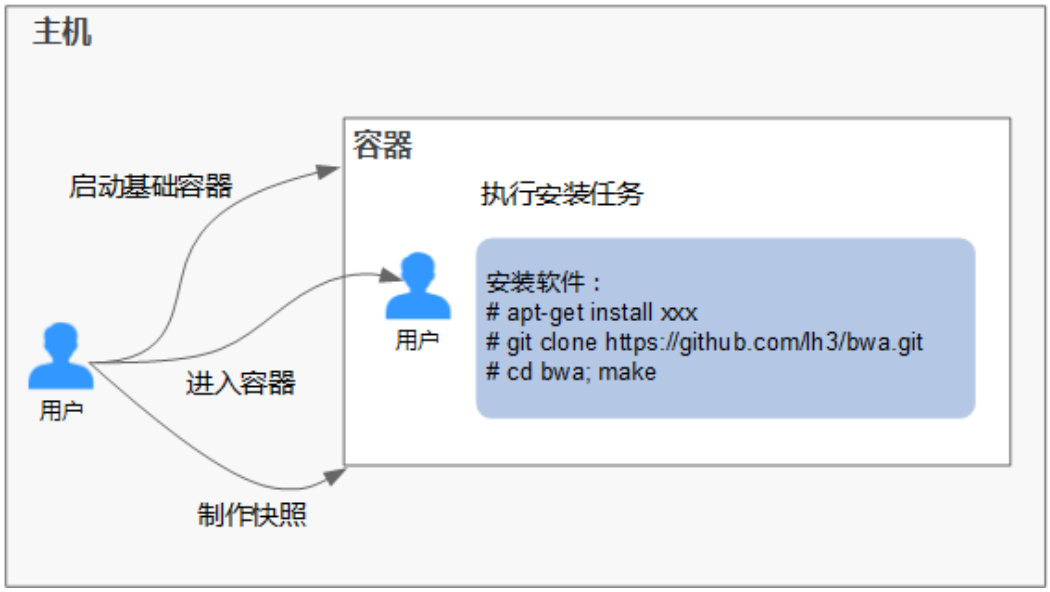
6.2.4 如何制作 Docker 镜像？

自己制作Docker镜像，主要有两种方法：

- 制作快照方式获得镜像（偶尔制作的镜像）：在基础镜像上，比如Ubuntu，先登录镜像系统并安装Docker软件，然后整体制作快照。
- Dockerfile方式构建镜像（经常更新的镜像）：将软件安装的流程写成DockerFile，使用Docker build构建成Docker镜像。

方法一：制作快照方式获得镜像

如果后续镜像没有变化，可采用方法一制作镜像。



具体操作如下：

1. 找一台主机，安装Docker软件。
2. 启动一个空白的基础容器，并进入容器。

   例如：启动一个CentOS的容器。

   docker run -it centos
3. 执行安装任务。

   yum install XXX

   git clone https://github.com/lh3/bwa.git

   cd bwa;make

   说明

   请预先安装好Git，并检查本机是否有ssh key设置。
4. 输入**exit**退出容器。
5. 制作快照。

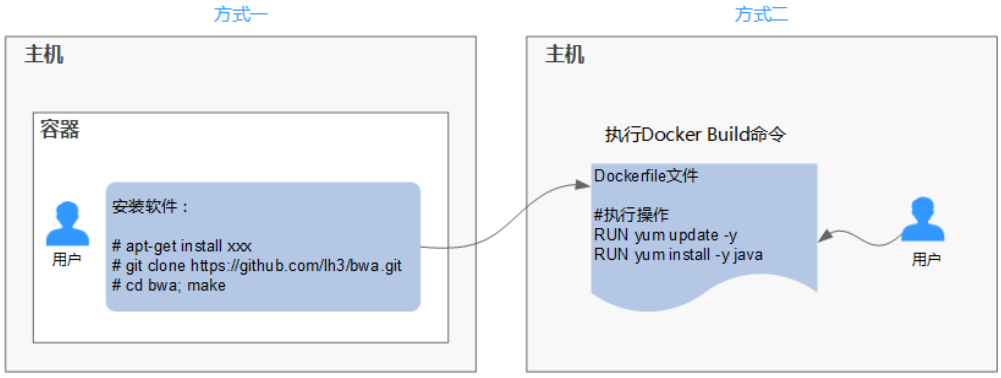
   docker commit -m "xx" -a "tsj" container-id tsj/image:tag

   - -a : 提交的镜像作者。
   - container-id : 操作步骤2中的容器id。可以使用 **docker ps -a** 查询得到容器id。
   - -m : 提交时的说明文字。
   - tsj/image:tag : 仓库名/镜像名:TAG名。
6. 执行**docker images**可以查看到制作完成的Docker镜像。

方法二：使用 Dockerfile 方式构建

如果后续镜像经常变更（例如某个软件更新版本），则需要采用方法二制作镜像。若仍采用方法一制作镜像，则每次变更都需要重新执行方法一的命令，操作过程比较繁琐，所以建议使用自动化制作镜像的方法。

其实就是将方法一制作镜像的方法，用文件方式写出来（文件名为DockerFile）。然后执行：**docker build -t tsj/image:tag.**命令（命令中“.”表示DockerFile文件的路径），自动完成镜像制作。



简单的DockerFile示例：

说明

如果依赖外部网络，请搭建网络环境，并保证网络可用。

```
#Version 1.0.1
FROM centos:latest
MAINTAINER ***u "***u@huaweicloud.com"

#设置root用户为后续命令的执行者
USER root

#执行操作
RUN yum update -y
RUN yum install -y java

#使用&&拼接命令
RUN touch test.txt && echo "abc" >>abc.txt

#对外暴露端口
EXPOSE 80 8080 1038

#添加网络文件
ADD https://www.baidu.com/img/bd_logo1.png /opt/

#设置环境变量
ENV WEBAPP_PORT=9090

#设置工作目录
WORKDIR /opt/

#设置启动命令
ENTRYPOINT ["ls"]

#设置启动参数
CMD ["-a", "-l"]

#设置卷
VOLUME ["/data", "/var/www"]

#设置子镜像的触发操作
ONBUILD ADD . /app/src
ONBUILD RUN echo "on build excuted" >> onbuild.txt
```

详细的操作步骤可以参考：《容器镜像服务 最佳实践》。

DockerFile 基本语法

- FROM：

  指定待扩展的父级镜像（基础镜像）。除注释之外，文件开头必须是一个FROM指令，后面的指令便在这个父级镜像的环境中运行，直到遇到下一个FROM指令。通过添加多个FROM命令，可以在同一个Dockerefile文件中创建多个镜像。

- MAINTAINER：

  声明创建镜像的作者信息：用户名、邮箱，非必须参数。

- RUN：

  修改镜像的命令，常用来安装库、安装程序以及配置程序。一条RUN指令执行完毕后，会在当前镜像上创建一个新的镜像层，接下来对的指令会在新的镜像上继续执行。RUN 语句有两种形式：

  - **RUN yum update**：在**/bin/sh**路径中执行的指令命令。
  - **RUN ["yum", "update"]**：直接使用系统调用**exec**来执行。
  - **RUN yum update && yum install nginx**：使用&&符号将多条命令连接在同一条RUN语句中。

- EXPOSE：

  指明容器内进程对外开放的端口，多个端口之间使用空格隔开。

  运行容器时，通过设置参数**-P**（大写）即可将EXPOSE里所指定的端口映射到主机上其他的随机端口，其他容器或主机可以通过映射后的端口与此容器通信。

  您也可以通过设置参数**-p**（小写）将Dockerfile中EXPOSE中没有列出的端口设置成公开。

- ADD：

  向新镜像中添加文件，这个文件可以是一个主机文件，也可以是一个网络文件，也可以使一个文件夹。

  - 第一个参数：源文件（夹）。
    - 如果是相对路径，必须是相对于Dockerfile所在目录的相对路径。
    - 如果是URL，会将文件先下载下来，然后再添加到镜像里。
  - 第二个参数：目标路径。
    - 如果源文件是主机上的zip或者tar形式的压缩文件，Docker会先解压缩，然后将文件添加到镜像的指定位置。
    - 如果源文件是一个通过URL指定的网络压缩文件，则不会解压。

- VOLUME：

  在镜像里创建一个指定路径(文件或文件夹)的挂载点，这个容器可以来自主机或者其它容器。多个容器可以通过同一个挂载点共享数据，即便其中一个容器已经停止，挂载点也仍可以访问。

- WORKDIR：

  为接下来执行的指令指定一个新的工作目录，这个目录可以是绝对目录，也可以是相对目录。根据需要，WORKDIR可以被多次指定。当启动一个容器时，最后一条WORKDIR指令所指的目录将作为容器运行的当前工作目录。

- ENV：

  设置容器运行的环境变量。在运行容器的时候，通过设置**-e**参数可以修改这个环境变量值，也可以添加新的环境变量。

  例如：

  docker run -e WEBAPP_PORT=8000 -e WEBAPP_HOST=www.example.com ...

- CMD：

  用来设置启动容器时默认运行的命令。

- ENTRYPOINT：

  用来指定容器启动时的默认运行的命令。区别在于：运行容器时添加在镜像之后的参数，对ENTRYPOINT是拼接，CMD是覆盖。

- 若在DockerFile中指定了容器启动时的默认运行命令为**ls -l**，则运行容器时默认启动命令为 **ls -l**，例如：
  - **ENTRYPOINT ["ls", "-l"]**：指定容器启动时的程序及参数为 **ls -l** 。
  - **docker run centos**：当运行centos容器时，默认执行的命令是**docker run centos ls -l**
  - **docker run centos -a**：当运行centos容器时拼接了-a参数，则默认运行的命令是**docker run centos ls -l -a**
- 若在DockerFile中指定了指定了容器启动时的默认运行命令为**--entrypoint**，则在运行容器时若需要替换默认运行命令，可以通过添加**--entrypoint**参数来替换Dockerfile中的指定。例如：

  docker run gutianlangyu/test --entrypoint echo "hello world"

- USER：

  为容器的运行及RUN、CMD、ENTRYPOINT等指令的运行指定用户或UID。

- ONBUILD：

  触发器指令。构建镜像时，Docker的镜像构建器会将所有的ONBUILD指令指定的命令保存到镜像的元数据中，这些命令在当前镜像的构建过程中并不会执行。只有新的镜像使用FROM指令指定父镜像为当前镜像时，才会触发执行。

  使用FROM以这个Dockerfile构建出的镜像为父镜像，构建子镜像时：

  ONBUILD ADD . /app/src：自动执行**ADD . /app/src**

