Cromwell 引擎使用指南

Cromwell 引擎使用指南

文档版本 01

发布日期 2020-12-17

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： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

1 什么是 Cromwell 引擎

Cromwell 是 Broad Institute 开发的工作流管理系统。通过 Cromwell 可以将 WDL（ Workflow Description Language ）描述的 workflow 运行在CCI容器中。您只需为作业运行时实际消耗的计算和存储资源付费，不需要支付资源之外的附加费用。本文将介绍如何使用在基因容器服务中使用Cromwell。

2 管理 Cromwell 引擎

创建 Cromwell 引擎

步骤1 登录**GCS控制台**，选择左侧导航栏的“环境管理”，在右侧页面单击“创建环境”。

步骤2 设置“默认环境”：是/否。若当前没有环境，则将要创建的环境即为默认环境。默认环境有且只有一个。在有多个集群时，执行测序任务时如果不指定投递环境，则将任务投递至该默认环境。

步骤3 选择“环境类型”为“Cromwell引擎”。

步骤4 选择“关联OBS存储”：OBS存储用于存储分析前后产生的数据，包括原始基因数据、流程执行中间数据及执行结果数据。

- 如果您已有可用桶，在创建环境中，选择对应的桶即可。
- 如果没有可用桶或是需要新建桶，请单击“创建OBS存储”创建。

说明

关联OBS存储功能目前正在公测中，如果您还未申请公测，请申请**CCI容器实例挂载方式使用OBS文件桶**。

步骤5 选择“命名空间”。

如果您在CCI中没有可用的命名空间，或不想使用已有命名空间，请单击“创建命名空间”，创建命名空间步骤请参见**命名空间**。

步骤6 上传“访问密钥”。

单击 单击上传，在弹出的对话框中上传已下载的访问密钥（AK/SK），单击“确认”。若没有访问密钥，请前往“我的凭证” 的**管理访问密钥**页面新增并下载访问密钥。

步骤7 选择“高速共享存储”。

“高速共享存储”对应的是文件存储服务SFS，用于存储流程中间数据。

如果没有可用的高速共享存储，或不想使用已有高速共享存储，请单击“创建存储”，在弹出的窗口中选择已有的文件存储，单击“导入”；如果没有文件存储，请单击“创建文件存储”，具体步骤请参见**创建SFS文件系统**。

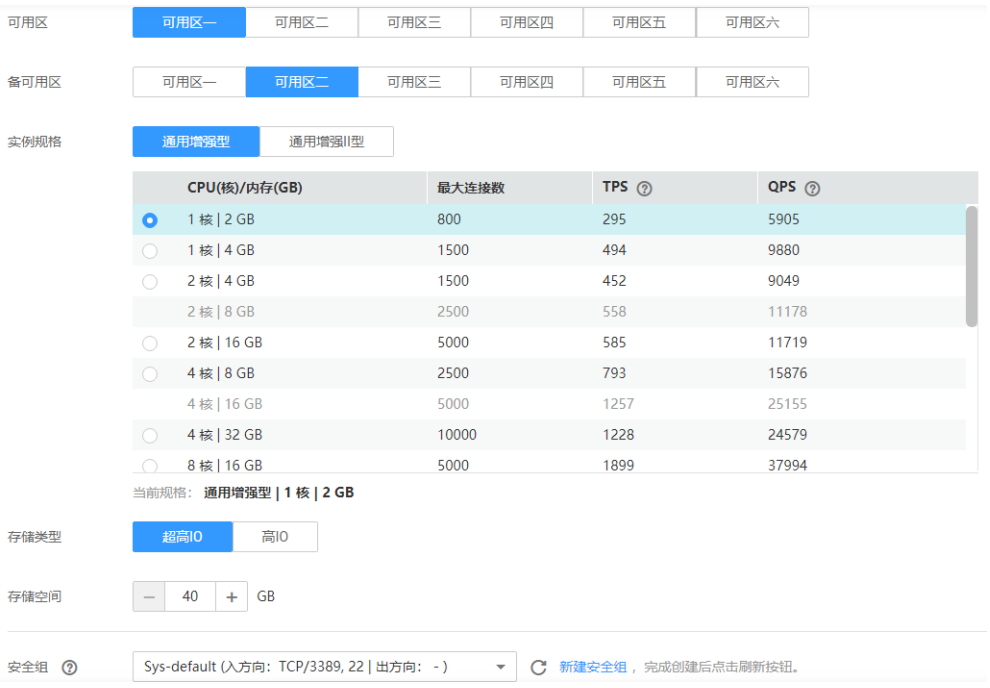
步骤8 配置RDS实例。

Cromwell引擎需要使用一个数据库存储数据。

您可以使用已有的RDS实例，也可以选择新建RDS实例，支持的mysql版本为5.6和5.7，新建RDS请填写如图2-1和图2-2所示参数。

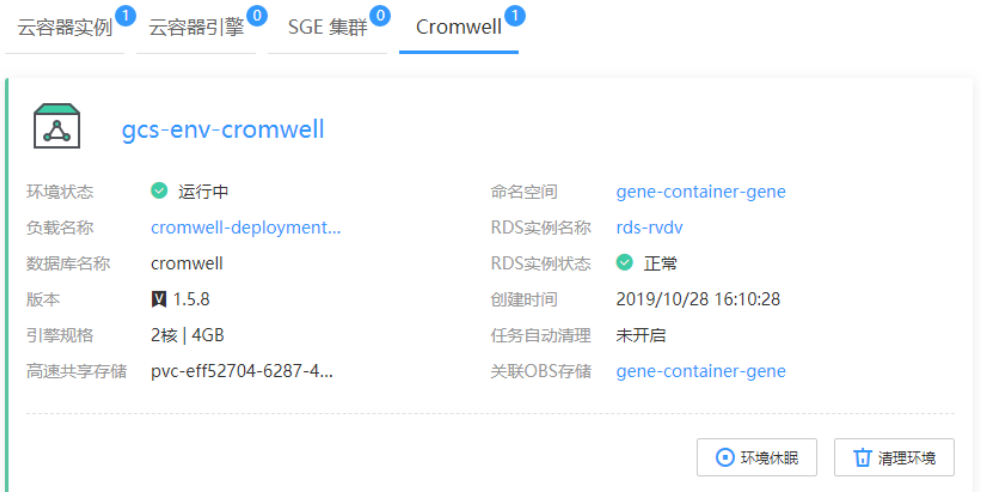
图 2-1 创建 RDS 实例

图 2-2 选择 RDS 实例规格



步骤9 单击“下一步”，填写“环境名称”，确认环境信息配置后单击“提交”。单击“环境管理列表”将跳转“环境管理”页面，环境状态为“运行中”，环境已创建成功。

图 2-3 Cromwell 环境



----结束

3 投递 Cromwell 任务

投递方法

当前Cromwell引擎可以通过命令行和SDK两种方式使用。

- SDK的使用方法请参见**Python SDK参考**。初始化后调用Cromwell相关接口即可投递任务。
- 命令行使用方法请参见**命令参考**。安装命令行工具后，使用**gcs sub wdl**即可投递Cromwell任务。

WDL 描述文件

Cromwell引擎使用WDL文件描述任务执行流程。WDL的语法规则请参见**1.0 specification**，更多WDL信息请参见**openwdl**。

常见配置说明

- **runtime** --- 运行时配置

  您可以在**WDL描述文件**中配置runtime参数，指定流程的运行时参数，示例如下：

```
runtime {
    docker: "swr.cn-north-4.myhuaweicloud.com/cromwell/gatk:4.1.0.0"
    cpu: "1"
    memory: "2G"
    disks: "/some/mnt 100 SSD"
}
```

表 3-1 runtime 参数说明

参数	必选	默认值	说明
docker	是	--	Task执行使用的镜像地址
cpu	否	1	Task执行需要的cpu核数，请根据实际情况选择
memory	否	“2G”	Task执行需要的内存大小，请根据实际情况选择

参数	必选	默认值	说明
disks	否	--	Task执行需要的本地磁盘规格和挂载点，其中规格支持SSD、SATA、SAS
maxRetries	否	3	Task执行失败后最大重试次数
continueOnReturnCode	否	0	当Task执行返回码不为指定值时，则认为task执行失败。当指定为true时，则认为所有返回码均为成功
failOnStderr	否	false	当标准输出流中检测到错误信息时，是否认为task执行失败

- **CallCaching** --- 缓存配置

  Cromwell能够检测到过去何时运行过作业，从而不必重新计算结果，节省运行时间和成本。Cromwell在之前运行的作业的缓存中搜索具有完全相同的命令和完全相同的输入的作业。如果在缓存中找到之前运行的作业，则使用之前作业的结果，而不是重新运行它。

  华为云Cromwell默认启用callCaching功能，您也可以在**options文件**中配置读写cache的开关，并在提交流程时进行指定。配置示例如下：

  ```
  {
     "write_to_cache": true, // 是否将执行结果写入缓存
     "read_from_cache": false // 是否从缓存中读取执行结果
  }
  ```

- **Filesystem** --- 对象存储配置

  Cromwell支持使用华为云对象存储服务（OBS）作为数据的输入和输出。您可将流程需要的样本数据的文件存放在OBS中，并在流程**inputs文件**中通过OBS的存放地址进行访问。示例配置如下：

  ```
  {
  "PreProcessingForVariantDiscovery_GATK4.flowcell_unmapped_bams": [
  "obs://NA12878_24RG/HJYFJ.4.NA12878.downsampled.query.sorted.unmapped.bam",
  "obs:// NA12878_24RG/HK3T5.8.NA12878.interval.filtered.query.sorted.unmapped.bam"
  ]
  }
  ```

4 查看执行结果

分析任务的执行时间较长，一般需要数小时，详细的时长与环境资源类型、环境资源大小、处理数据大小等相关。您可以通过**GCS控制台**左侧导航栏的“执行结果”查看执行结果或是操作任务。

- 查看执行结果，您可以实时查看分析任务的执行过程、数据、状态、结果等信息，请参见**查看执行结果**。
- 管理执行结果，您可以根据需求对任务执行删除、备份操作。备份操作请参见**备份执行结果**，其它操作请参见**管理执行结果**。

执行状态说明

从执行状态可以确定当前任务所属阶段，已有执行状态如下。

- 初始化：任务准备执行中。
- 执行中：任务正在执行中，可更新任务优先级、查看任务日志、删除任务、停止任务。
- 成功：任务执行成功，可查看任务日志、删除任务。
- 失败：任务执行失败，可查看任务日志、删除任务、重试执行任务。
- 已停止：任务已停止，可修改任务优先级、查看任务日志、删除任务、启动任务、更新流程的配置参数。
- 停止中：任务正在停止中。
- 删除中：任务正在删除中。

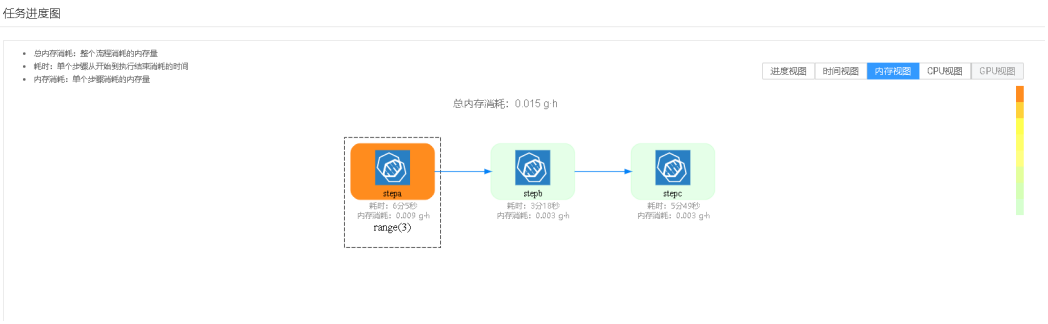
查看执行结果

执行结果页面可以查看所有任务，任务按创建时间排序，最新的排在最前面。

对每一个任务，您可以实时查看分析任务的执行过程（包括执行结果热力图）、数据、状态、结果等信息。

图 4-1 执行结果

图 4-2 热力图



管理执行结果

任务提交后，除查看执行结果外，您可以在执行结果页面对任务做操作。

- 删除任务：删除任务执行结果，所有任务均可删除，删除后不可恢复，请谨慎操作。
  - 删除一个任务。在执行结果列表，单击“操作”列的“删除”（上图中的7），删除任务。
  - 批量删除任务。在执行结果列表中的“任务名称”勾选需要批量删除的任务，单击“删除”（上图中的8），删除任务。

须知

删除任务仅能删除任务在GCS中的记录，并不会删除任务的存储在OBS或是SFS中的原始数据或是中间数据，如需删除，请手动删除。

备份执行结果

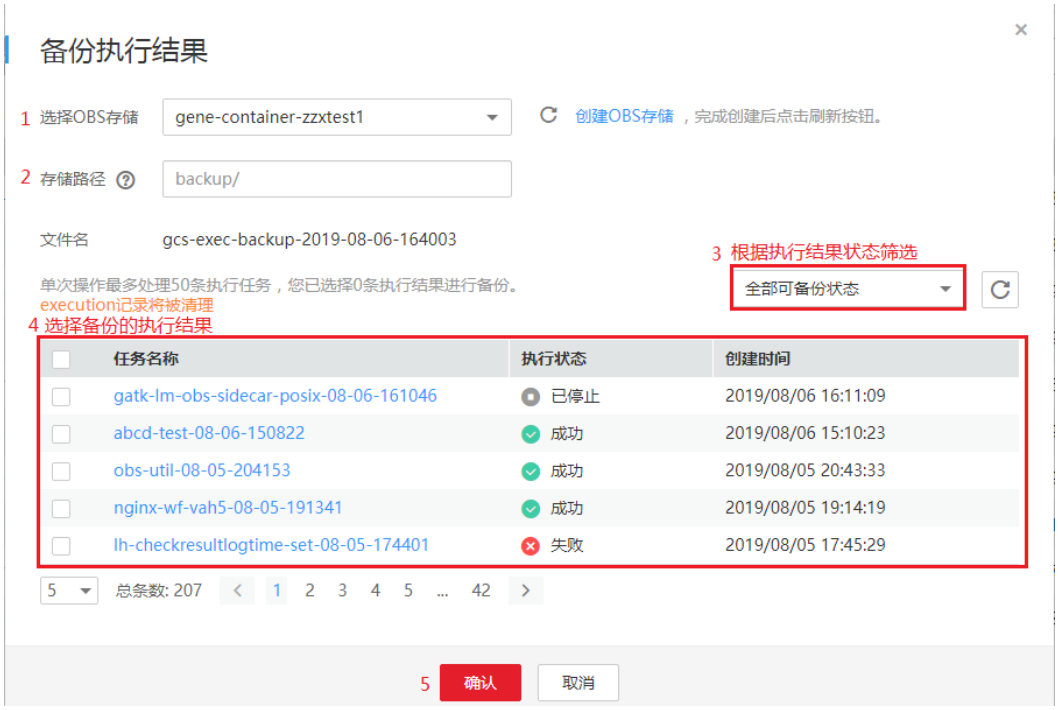
备份后的执行结果数据，将以CSV文件格式保存在指定OBS存储的指定路径下以CSV文件的方式，并从“执行结果”列表中删除。为保证新的分析任务可以正常执行，建议定期备份历史数据。

步骤1 登录GCS控制台，选择左侧导航栏的“执行结果”，在右侧页面单击“备份”。

图 4-3 备份

步骤2 在弹出的容器中填写参数，并勾选需要备份的任务，完成后单击“确认”。

图 4-4 备份执行结果



- 选择OBS存储。选择待备份执行结果存储的OBS存储。可在下拉框中选择已有存储，或是创建新的OBS存储。
- 存储路径。待备份执行结果在OBS存储中的存储路径，建议存储在“backup/”路径下。斜杠（/）表示分隔并创建多层级文件夹。
- 文件名。系统指定，无法修改。备份完成后，可以通过该文件名在存储路径中找到备份的CSV文件。
- 选择通过执行结果状态过滤执行结果（上图中的3），从而快速备份特定状态的执行结果。

步骤3 备份成功后，界面将提示备份成功，您可以通过界面提示链接“数据存储列表”，或是在“数据管理”>“私有数据”查看备份结果。

通过**步骤2**中的文件名找到备份文件，单击“操作”中的“下载”，即可查看备份的执行结果数据。

----结束

5 示例

本节通过一个简单的示例说明Cromwell的使用方法。

步骤 1：WDL 流程文件编写

使用Cromwell首先编写WDL流程文件，定义任务如何工作。Cromwell引擎使用WDL文件描述任务执行流程，WDL的语法规则请参见**1.0 specification**。

下面是一个简单示例WDL流程示例“example.wdl”。详细解释如下：

- 这个流程一共有3个步骤stepa、stepb、stepc，首先执行stepa，stepa使用firstInput 作为输入，且指定并发数为3；然后执行stepb，stepb使用stepa的输出stepa.out作为输入；然后执行stepc，stepc使用stepb的输出stepb.out作为输入。
- 最后将stepc.out作为最终输出结果。
- stepa、stepb、stepc中都定义了runtime运行时，指定了使用“swr.cn-north-1.myhuaweicloud.com/op_svc_gcs_container/cromwell:1.5.8”这个镜像，且指定了使用的资源大小为1核2G。

```
workflow example{
  File firstInput
  scatter (idx in range(3)) {
    call stepa { input: in=firstInput }
  }
  call stepb { input: in=stepa.out }
  call stepc { input: in=stepb.out }
  output {
    File result = stepc.out
  }
}
task stepa {
  File in
  command { cat ${in} > outputa.txt && cat outputa.txt}
  output { File out = "outputa.txt" }
  runtime {
    docker: "swr.cn-north-1.myhuaweicloud.com/op_svc_gcs_container/cromwell:1.5.8"
    cpu: "1"
    memory: "2G"
  }
}
task stepb {
  Array[File] in
  command { cat ${write_lines(in)} > outputb.txt && cat outputb.txt}
  output { File out = "outputb.txt" }
  runtime {
    docker: "swr.cn-north-1.myhuaweicloud.com/op_svc_gcs_container/cromwell:1.5.8"
    cpu: "1"
```



```
    memory: "2G"
  }
}
task stepc {
  File in
  command { cat ${in} > outputc.txt && cat outputc.txt}
  output { File out = "outputc.txt" }
  runtime {
    docker: "swr.cn-north-1.myhuaweicloud.com/op_svc_gcs_container/cromwell:1.5.8"
    cpu: "1"
    memory: "2G"
  }
}
```

上面流程定义了一个输入文件File firstInput，输入文件使用“.inputs”文件定义，如下“example.inputs”文件所示，example.firstInput使用“obs://gcs-tool-cn-north-1/example.txt”这个文件作为输入。

```
{
  "example.firstInput": "obs://gcs-tool-cn-north-1/example.txt",
}
```

步骤 2：准备 Cromwell 环境

- 定义了流程后，需要准备Cromwell环境，Cromwell环境创建方法请参见**创建Cromwell引擎**。
- example.wdl使用了“swr.cn-north-1.myhuaweicloud.com/op_svc_gcs_container/cromwell:1.5.8”这个镜像，这个镜像是公共镜像，可以直接使用。在其他流程中使用镜像，您需要先制作好镜像，然后上传到华为云容器镜像服务中，然后再使用。
- example.inputs使用了OBS桶中的文件，“obs://gcs-tool-cn-north-1/example.txt”是一个公共可读文件，可以直接使用。在其他流程中使用文件，您可以先上传到OBS中，然后再使用。

步骤 3：投递任务

环境准备好后，就可以投递任务了。

当前Cromwell引擎可以通过命令行和SDK两种方式使用。

- SDK的使用方法请参见**Python SDK参考**。初始化后调用Cromwell相关接口即可投递任务。
- 命令行使用方法请参见**命令参考**。安装命令行工具后，使用**gcs sub wdl**即可投递Cromwell任务。

这里使用命令行工具作为示例投递任务，如下所示。

```
gcs sub wdl example.wdl -i example.inputs -s gcs-env-test
```

gcs sub wdl是命令，example.wdl是流程文件，-example.inputs为输入文件，gcs-env-test是**步骤2：准备Cromwell环境**中创建的Cromwell环境的名称。

回显内容如下。

```
current environment is gcs-env-test
create wdl succeed
{
"id":
"aa215e0b-c896-4138-865e-6cf7921246b6",
"status": "Submitted",
"message": ""
}
```


步骤 4：查看执行结果

任务投递后，您可以在基因容器控制台中实时查看任务的执行过程（包括执行结果热力图）、数据、状态、结果等信息。您还可以使用命令行直接查看，具体请参见**gcs get wdl**。

图 5-1 执行结果