WDL 101: Running WDLs using Cromwell

October, 2022

Table of Contents

# About this Course

This course is intended to be an introduction for users at the Fred Hutch to using our pre-configured Cromwell resources to run WDL workflows using our SLURM cluster.

# 1 Introduction

WDL is a open specification for a workflow description language that originated at the Broad but has grown to a much wider audience over time. WDL workflows can be run using an engine, which is software that interprets and runs your WDL on various high performance computing resources, such as SLURM (the Fred Hutch local cluster), AWS, Google and Azure. At the Fred Hutch we have configured a software from the Broad called Cromwell to allow us to run WDLs on our local cluster that then can be easily ported to other cloud based compute infrastructure when desired. This allows us to simplify our workflow testing and design, leverage WDL for smaller scale work that does not need the cloud, and can let users of all kinds manage their workflow work over time via this tool.

## 1.1 Pre-Requisites

### 1.1.1 rhino Access

Either you need to have gotten access to use the Fred Hutch SLURM cluster prior OR have taken our “Cluster 101” course. You can find this **free** course [here](https://leanpub.com/courses/fredhutch/fredhutchcluster101). To follow this course, you’ll need to know how to get onto the rhino nodes. If this is not something you are familiar with, you also may want to read a bit more about the use of our cluster over at [SciWiki](https://sciwiki.fredhutch.org/) in the Scientific Computing section about Access Methods, and Technologies.

### 1.1.2 AWS Credentials

If you want to run workflows on data stored in AWS S3, you’ll need to have set up your AWS credentials first.  
As of version 1.3 of the diy-cromwell-server configuration, if you have credentials, then the Cromwell server will be configured to allow input files to directly specified using their AWS S3 url.  
However if you do not have AWS credentials or aren’t using data stored in AWS S3, then you don’t have to do anything.

## 1.2 What is Cromwell?

Cromwell is a workflow engine (sometimes called a workflow manager) software developed by the Broad which manages the individual tasks involved in multi-step workflows, tracks job metadata, provides an API interface and allows users to manage multiple workflows simultaneously.  
Cromwell isn’t the only WDL “engine” that exists, but it is a tool that has been configured for use on the Fred Hutch gizmo cluster in order to make running workflows here very simple.

## 1.3 Getting Started with Cromwell

### 1.3.1 Database Setup

One of the strengths of using Cromwell is that it can allow you to keep track of what you’ve already done so you don’t have to re-compute jobs that have already completed successfully even if some of your workflow fails. To leverage this feature, called “call caching” and also use our computing resources efficiently, our configuration of Cromwell will help you set up a Cromwell server for 7 days at a time and creates a database that persists forever that the Cromwell servers can communicate with. We have found that by using a MySQL database for your Cromwell server to store information in, it will run faster and be better able to handle simultaneous workflows while also making all the metadata available to you during and after the run.

Thus for each user, we suggest setting up a database which only needs to be done one time per user using these steps:

1. Go to [DB4Sci](https://mydb.fredhutch.org/login) and see the Wiki entry for DB4Sci [here](https://sciwiki.fredhutch.org/scicomputing/store_databases/#db4sci--previously-mydb).
2. Login using Fred Hutch credentials, choose Create DB Container, and choose the MariaDB option.
3. The default database container values are typically fine, EXCEPT you likely need either weekly or no backups (no backups preferred) for this database.
4. Save the DB/Container Name, DB Username and DB Password as you will need them for the configuration step.

Once you click submit, a confirmation screen will appear (hopefully), and you’ll need to note which Port is specified. This is a 5 digit number currently.

Now you have a “container” in which to run a database, but the database itself does not yet exist.  
You need to create an empty database in that container and then Cromwell will do the rest the first time you start up a server.

To create the database, you’ll need to go to Terminal and connect to rhino then follow these instructions:

ml MariaDB/10.5.1-foss-2019b  
mysql --host mydb --port <Port> --user <username> --password

It will then prompt you to enter the DB password you specified during setup. Once you are are a “mysql>” prompt, you can do the following. >Note, we suggest you name the database inside the container the same as the container, but you cannot include dashes in your database name. In the future, DB4Sci may also set up the database inside the container for you, in which case you would be provided a database name as well during setup.

MariaDB [(none)]> create database <DB Name>;  
# It should do its magic  
MariaDB [(none)]> exit

Now you’re ready to go and never have to set up the database part again and you can use this database to manage all your work over time!

### 1.3.2 Start up a Cromwell server

#### 1.3.2.1 Customize Your Configuration

To start up your first Crowmell server job, you first need to decide where you want to keep your Cromwell configuration files.  
This must be a place where rhino can access them, such as in your Home directory, which is typically the default directory when you connect to the rhinos.  
We suggest you create a cromwell-home folder (or whatever you want to call it) and follow these git instructions to clone it directly.

Then you will set up the customizations that you’re going to want for your server(s) by making user configuration file(s) in your cromwell-home or wherever you find convenient.

You can manage mulitple Cromwell profiles this way by just maintaining different files full of credentials and configuration variables that you want.

To get started, do the following on rhino:

mkdir -p cromwell-home  
cd cromwell-home  
git clone --branch main https://github.com/FredHutch/diy-cromwell-server.git  
  
  
cp ./diy-cromwell-server/cromUserConfig.txt .  
## When you are first setting up Cromwell, you'll need to put all of your User Customizations into this `cromUserConfig.txt` which can serve as a template.   
## After you've done this once, you just need to keep the path to the file(s) handy for the future.

In cromUserConfig.txt there are some variables that allow users to share a similar configuration file but tailor the particular behavior of their Cromwell server to best suit them.  
The following text is also in this repo but these are the customizations you’ll need to decide on for your server.

################## WORKING DIRECTORY AND PATH CUSTOMIZATIONS ###################  
## Where do you want the working directory to be for Cromwell (note: this process will create a subdirectory here called "cromwell-executions")?   
### Suggestion: /fh/scratch/delete90/pilastname\_f/username/  
SCRATCHDIR=/fh/scratch/delete90/...  
  
## Where do you want logs about individual workflows (not jobs) to be written?  
## Note: this is a default for the server and can be overwritten for a given workflow in workflow-options.  
### Suggestion: /fh/fast/pilastname\_f/cromwell/workflow-logs  
WORKFLOWLOGDIR=~/cromwell-home/workflow-logs  
  
## Where do you want to save Cromwell server logs for troubleshooting Cromwell itself?  
### Suggestion: ~/home/username/cromwell-home/server-logs  
SERVERLOGDIR=~./cromwell-home/server-logs  
  
################ DATABASE CUSTOMIZATIONS #################  
## DB4Sci MariaDB details (remove < and >, and use unquoted text):  
  
CROMWELLDBPORT=...  
CROMWELLDBNAME=...  
CROMWELLDBUSERNAME=...  
CROMWELLDBPASSWORD=...  
  
## Number of cores for your Cromwell server itself - usually 4 is sufficient.   
###Increase if you want to run many, complex workflows simultaneously or notice your server is slowing down.  
NCORES=4  
  
## Length of time you want the server to run for.   
### Note: when servers go down, all jobs they'd sent will continue. When you start up a server the next time  
### using the same database, the new server will pick up whereever the previous workflows left off. "7-0" is 7 days, zero hours.  
SERVERTIME="7-0"

Note: For this server, you will want multiple cores to allow it to multi-task. Memory is less important when you use an external database. If you notice issues, the particular resource request for the server job itself might be a good place to start adjusting, in conjunction with some guidance from SciComp or the Slack [Question and Answer channel](https://fhbig.slack.com/archives/CD3HGJHJT) folks.

#### 1.3.2.2 Kick off your Cromwell server

Now that you’ve configured your future Cromwell servers, you can kick off your first Cromwell server job. Go to rhino to your “cromwell-home” anddo the following:

## You'll want to put `cromwell.sh` somewhere handy for future use, we suggest:  
cp ./diy-cromwell-server/cromwell.sh .  
  
## Then you'll want to make the script "executable":  
chmod +x cromwell.sh  
  
# Then simply start up Cromwell by executing the script and passing it the path to your configuration file.   
./cromwell.sh cromUserConfig.txt

Much like the grabnode command you may have used previously, the script will run and print back to the console instructions once the resources have been provisioned for the server. You should see something like this:

Your configuration details have been found...  
Getting an updated copy of Cromwell configs from GitHub...  
Setting up all required directories...  
Detecting existence of AWS credentials...  
Credentials found, setting appropriate configuration...  
Requesting resources from SLURM for your server...  
Submitted batch job 50205062  
Your Cromwell server is attempting to start up on \*\*node/port gizmok30:2020\*\*. If you encounter errors, you may want to check your server logs at /home/username/cromwell-home/server-logs to see if Cromwell was unable to start up.  
Go have fun now.

NOTE: Please write down the node and port it specifies here. This is the only place where you will be able to find the particular node/port for this instance of your Cromwell server, and you’ll need that to be able to send jobs to the Crowmell server. If you forget it, scancel the Cromwell server job and start a new one.

While your server will normally stop after 7 days (the default), at which point if you have jobs still running you can simply restart your server and it will reconnect to existing jobs/workflows. However, if you need to take down your server for whatever reason before that point, you can go to rhino and do:

# Here `username` is your Fred Hutch username  
squeue -u username  
## Or if you want to get fancy:  
squeue -o '%.18i %.9P %j %.8T %.10M %.9l %.6C %R' -u username  
  
## You'll see a jobname "cromwellServer". Next to that will be a JOBID. In this example the JOBID of the server is 50062886.  
  
scancel 50062886

Congrats you’ve started your first Cromwell server!! Now on to how to submit a WDL workflow to it.

# 2 Using Shiny to Manage Workflows

Now that you’ve configured your first Cromwell sever, let’s submit some test workflows to it using the Fred Hutch Shiny app!

Note: especially the first time you set up a Cromwell server, it will be busy for a few minutes setting up the database and doing all the work behind the scenes for you. Once it’s “ready” to listen for workflows it will start “listening” for instructions via the Shiny app (or other methods we’ll discuss later in the course). It may take 2-3 minutes before you can follow the rest of these instructions the first time. The time it takes is much shorter in the future (more like ~1 minute).

## 2.1 Cromwell app

You can find our Fred Hutch Shiny app here: <https://cromwellapp.fredhutch.org/>

This shiny app will let you use a graphic interface to submit and manage workflows you’ve written in WDL.

### 2.1.1 Login

While this Shiny app runs all the time, in order for it to know where to look for your particular information, you’ll need to to “login” by clicking the “Connect to Server” button on the left.

When you click login, a box will appear where you will input your node:port combo from the output of the Cromwell server kickoff process before (it will be in the format of “gizmob5:39071”).

If your server is not yet fully ready to listen for workflows you may see this error result:

If so, just wait 1-2 more minutes (if it’s the first time you’ve set up a server, or less if it’s a future instance) and try again. Once the Shiny app can talk to your sever, you’ll see this result screen:

### 2.1.2 Submit Jobs Tab

Once you’ve connected your server to the Shiny app, you can start by using the “Submit Jobs” tab on the left.

Here you’ll see a series of sections that will allow you to do several things.

#### 2.1.2.1 Validate a workflow

This checks the format of your workflow files to make sure you have a valid file in a known format that Cromwell can iterpret. It does not perform a “dry run” or check to see if any of your inputs are actually available, only that it can interpret what you told it.

#### 2.1.2.2 Submit a workflow

This will let you upload the files that contain your workflow description (a WDL), and up to two different sets of input lists (in JSON form). You can run a workflow with no input JSON, one input JSON, or two input JSONs (which will be concatenated or the second will overwrite the first if the same variable is declared in both). You can upload a workflow options JSON (which you’ll learn about in future classes), as well as providing text labels of your choosing to workflows if you’d like.

#### 2.1.2.3 Abort a workflow

Sometimes you realize you might just want to kill a workflow. Using the workflow submission id, you can specifically kill workflows by using this box. Note it will take Cromwell some time to coordinate SLURM job cancellations but it will clean everything up for you, it’s just not instant.

#### 2.1.2.4 Troubleshoot a workflow

Especially in the beginning if you have catastrophic workflow failures and you can’t even figure out what’s going on, you can come back to this Troubleshoot box to retreive the entire, unformatted JSON output of all metadata Cromwell has about your workflow here. You probably are better served by the next tab in the app for checking up on how your workflow is going, but if there’s nothing there that’s helpful, then this box is where you’ll want to go.  
> Note: this is not for the faint of heart, but it will give you hints once you get used to understanding what Cromwell is telling you.

### 2.1.3 Track Jobs Tab

Once you’ve submitted a workflow, you’ll want to track how it’s going in the Track Jobs tab.

#### 2.1.3.1 Workflows of History

Here you’ll see that you can query for as many days of history worth of workflows that you’d like, filter that result (for instance if you have submitted a LOT of workflows and the app is slow) for workflows with a specific name or with specific status(es) (such as failed, suceeded, etc).

Then you’ll see a plot of all the workflows returned and how long they have run, as well as what their status is.

Then you’ll see a table of the metadata about of all the workflows returned. Click on the workflow you’re interested in to populate the rest of the tables.

#### 2.1.3.2 Diving into a Workflow

Once you’ve selected a workflow row, you’ll see some summary information about that workflow.

You can see a plot of the timing and outcomes of all the calls in that workflow.

Then there is a table of all of those calls in which you can find a variety of useful information such as the directory where the job is working (callRoot), the SLURM job id it had/has, what computing resources or software environment was used, and what it’s status is.

Then you can use the Job Failures nad Call Caching tables to retreive information relevant to those processes by clicking the “Get/Refresh … Metadata” buttons (sometimes these can be quite large, and thus they do not load until you want them).

Finally, once a workflow suceeds, Cromwell can tell you (and this Shiny app can help you download) a table of all the outputs to the workflow itself (note this is not every file created, only the ones you specify as the “results” of the workflow overall). This is useful to go find those results and interact with them, archive them, or otherwise copy them to longer term storage for use.

### 2.1.4 Run Test Workflows

Now that you know how to use the app, it’s time to run a test workflow.

We have curated a number of basic workflows that you can use to test to see if your Cromwell server is set up correctly and for you to test out how working with Cromwell is done. See our [Test Workflow folder](https://github.com/FredHutch/diy-cromwell-server/tree/main/testWorkflows) once your server is up and run through the tests specified in the markdown there.

NOTE: For those test workflows that use Docker containers, know that the first time you run them, you may notice that jobs aren’t being sent very quickly. That is because for our cluster, we need to convert those Docker containers to something that can be run by Singularity. The first time a Docker container is used, it must be converted, but in the future Cromwell will used the cached version of the Docker container and jobs will be submitted more quickly.

# 3 Cromwell at the Fred Hutch

Cromwell can help run WDL workflows on a variety of computing resources such as SLURM clusters, as well as AWS, Google and Azure cloud computing systems. Using WDL workflows allows users to focus on their workflow contents rather than the intricacies of the particular computing platforms they are using.  
However, there are always optimizations of how those workflows run taht may be somewhat specific to the various computing tools you may be using.  
We’ll discuss some of the available customizations to help you run WDLs on our cluster in a simple way that still allows those workflows to be portable to other computing platforms.

## 3.1 Standard Runtime Variables

These runtime variables are both the defaults for our Fred Hutch configuration and also standard runtime variables you can use on other computing platforms too.

* cpu: 1
  + An integer number of cpus you want for the task
* memory: 2000
  + An integer number of MB of memory you want to use for the task
* docker: "ubuntu:latest"
  + A specific Docker container to use for the task. For the custom Hutch configuration, docker containers can be specified and the necessary conversions (to Singularity) will be performed by Cromwell (not the user). Note: when docker is used, soft links cannot be used in our filesystem, so workflows using very large datasets may run slightly slower due to the need for Cromwell to copy files rather than link to them.

## 3.2 Fred Hutch Custom Runtime Variables

For the gizmo cluster, the following runtime variables are available that are customized to our configuration.  
What is specified below is the current default as written, you can edit these in the config file if you’d like OR you can specify these variables in your runtime block in each task to change only the variables you want to change from the default for that particular task.

* walltime: "18:00:00"
  + A string of date/time that specifies how many hours/days you want to request for the task
* partition: "campus-new"
  + Which partition you want to use, the default is campus-new but whatever is in the runtime block of your WDL will overrride this. Other options currently include restart or short but check [SciWiki](https://sciwiki.fredhutch.org/scicomputing/) for more updated information.
* modules: ""
  + A space-separated list of the environment modules you’d like to have loaded (in that order) prior to running the task.
* dockerSL: "ubuntu:latest"
  + This is a custom configuration for the Hutch that allows users to use docker and softlinks only to specific locations in Scratch. It is helpful when working with very large files.
* account: "paguirigan\_a"
  + This allows users who run jobs for multiple PI accounts to specify at the level of a task which account to use for a given job to manage cluster allocations.

### 3.2.1 Software environments

#### 3.2.1.1 Modules

At Fred Hutch we have huge array of pre-curated software modules installed on our SLURM cluster which you can [read about in SciWiki](https://sciwiki.fredhutch.org/scicomputing/compute_scientificSoftware/).  
The custom configuration of our Cromwell server software here at the Hutch allows users to specify modules (or combinations of modules by simply adding them on in a space separated string) available on the cluster to use for individual tasks in a workflow. The module(s) desired can be used for a task by specifying in the runtime block of your calls the following:

runtime {  
 modules: "GATK/4.2.6.1-GCCcore-11.2.0 SAMtools/1.16.1-GCC-11.2.0"  
}

In this example, the GATK module will be loaded first, followed by the SAMtools modiule. In this example you’ll note the “toolchain” used to build both modules are the same (“GCC-11.2.0”). It is important to ensure when you load modules together for a single task that they are compatible with each other.

#### 3.2.1.2 Docker

However, if you want to move your WDL workflow to the cloud in the future, you’ll want to leverage Cromwell’s pre-configured ability to run your tasks on gizmo in Docker containers. This configuration allows users to specify docker containers in their runtime blocks, allows Cromwell to maintain a local cache of previously used containers, and facilitates the pull of Docker containers and conversion for use. This behavior allows us to evade rate-limiting by DockerHub and improves speed of your workflows.

#### 3.2.1.3 Find Support

If you’re unfamiliar with resources for using Docker or to learn more about using modules and software environments see the “Guidance and Support” section of this course. Software management in workflows can be don in many ways so finding what works best for your work is often an iterative process.

## 3.3 Guidance and Support

There are a variety of resources on campus from the [SciWiki](https://sciwiki.fredutch.org) to SciComp office hours (found in the Research Data Support Teams team) to asking questions in the [FH Data Slack #question-and-answer channel](https://fhdata.slack.com/archives/CD3HGJHJT) to coming to the [Effective Computing Drop in hours held by Fred Hutch DaSL](https://fhdata.slack.com/archives/C03T4QPNDA9).

### 3.3.1 Effective Computing User Group

This is a drop-in style biweekly meeting that provides resources for folks to talk about their computational work and improve over time. Find current events by checking out the FH Data Slack workspace [here](https://fhdata.slack.com).

### 3.3.2 Slack Workflow Managers

In the [FH Data Slack workspace](https://fhdata.slack.com/) there is a specific channel where you can find help from peers and staff who support Cromwell and WDL here at the Hutch called [#workflow-managers](https://fhdata.slack.com/archives/CJFP1NYSZ)

# 4 Using the fh.wdlR Package

The Shiny app deployed at Fred Hutch to allow users to submit workflows to running Crowmell servers is built using an R package which is freely available in GitHub.

<https://github.com/FredHutch/fh.wdlR>

This R package allows you to use R/RStudio on your local machine (on VPN or on campus) to directly submit workflows to your server from the command line, and lets you track calls and workflow execution status directly.

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [Amy Paguirigan] |
| Content Editor(s)/Reviewer(s) | Checked your content |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value   
## version R version 4.0.2 (2020-06-22)  
## os Ubuntu 20.04.3 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2022-10-27   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.3)   
## bookdown 0.24 2022-02-15 [1] Github (rstudio/bookdown@88bc4ea)   
## callr 3.4.4 2020-09-07 [1] RSPM (R 4.0.2)   
## cli 2.0.2 2020-02-28 [1] RSPM (R 4.0.0)   
## crayon 1.3.4 2017-09-16 [1] RSPM (R 4.0.0)   
## desc 1.2.0 2018-05-01 [1] RSPM (R 4.0.3)   
## devtools 2.3.2 2020-09-18 [1] RSPM (R 4.0.3)   
## digest 0.6.25 2020-02-23 [1] RSPM (R 4.0.0)   
## ellipsis 0.3.1 2020-05-15 [1] RSPM (R 4.0.3)   
## evaluate 0.14 2019-05-28 [1] RSPM (R 4.0.3)   
## fansi 0.4.1 2020-01-08 [1] RSPM (R 4.0.0)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.6.1 2022-01-22 [1] CRAN (R 4.0.2)   
## htmltools 0.5.0 2020-06-16 [1] RSPM (R 4.0.1)   
## knitr 1.33 2022-02-15 [1] Github (yihui/knitr@a1052d1)   
## lifecycle 1.0.0 2021-02-15 [1] CRAN (R 4.0.2)   
## magrittr 2.0.2 2022-01-26 [1] CRAN (R 4.0.2)   
## memoise 1.1.0 2017-04-21 [1] RSPM (R 4.0.0)   
## pkgbuild 1.1.0 2020-07-13 [1] RSPM (R 4.0.2)   
## pkgload 1.1.0 2020-05-29 [1] RSPM (R 4.0.3)   
## prettyunits 1.1.1 2020-01-24 [1] RSPM (R 4.0.3)   
## processx 3.4.4 2020-09-03 [1] RSPM (R 4.0.2)   
## ps 1.3.4 2020-08-11 [1] RSPM (R 4.0.2)   
## purrr 0.3.4 2020-04-17 [1] RSPM (R 4.0.3)   
## R6 2.4.1 2019-11-12 [1] RSPM (R 4.0.0)   
## remotes 2.2.0 2020-07-21 [1] RSPM (R 4.0.3)   
## rlang 0.4.10 2022-02-15 [1] Github (r-lib/rlang@f0c9be5)   
## rmarkdown 2.10 2022-02-15 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.2 2020-11-15 [1] CRAN (R 4.0.2)   
## sessioninfo 1.1.1 2018-11-05 [1] RSPM (R 4.0.3)   
## stringi 1.5.3 2020-09-09 [1] RSPM (R 4.0.3)   
## stringr 1.4.0 2019-02-10 [1] RSPM (R 4.0.3)   
## testthat 3.0.1 2022-02-15 [1] Github (R-lib/testthat@e99155a)   
## usethis 2.1.5.9000 2022-02-15 [1] Github (r-lib/usethis@57b109a)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2022-02-15 [1] Github (yihui/xfun@74c2a66)   
## yaml 2.2.1 2020-02-01 [1] RSPM (R 4.0.3)   
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library

# 5 References