R jobs in the heavens

Running R jobs at large scale (remotely)

- A good point: Collaboration and sharing codes and analysis

- Local clusters... you may have access to one in you institution, but

  - high cost for maintenance and administration...
  - it must be accessible to all researchers... independently from subject
  - young guns and basic sciences are not very appealing to deserve access...

Cloud possibilities

Amazon Web Services - AWS

- Getting started with R on Amazon Web Services - https://aws.amazon.com/blogs/opensource/getting-started-with-r-on-amazon-web-services/
- Running an R Code on AWS Batch - https://medium.com/geekculture/running-an-r-container-on-aws-batch-on-production-9be336c34f95

  - Jobs are submitted via Docker files with help of `packrat R` package
  - cost? only free trial?

Azure - Microsoft

- R workloads on Azure Batch - https://azure.microsoft.com/pt-br/blog/r-workloads-on-azure-batch/
- Tutorial: Run a parallel workload with Azure Batch using the .NET API - https://docs.microsoft.com/en-us/azure/batch/tutorial-parallel-dotnet

  - paid if you have an institutional account?

Google Cloud Platform

- Running R at Scale on Compute Engine - https://cloud.google.com/architecture/running-r-at-scale
  - If not eligible for a free trial, "... a 6-node cluster composed of n1-standard-4 instances, would cost $1.84/hr ..."

Oracle cloud computing

- Using Oracle R Enterprise Embedded R Execution
  - free credits for new accounts... better performance? lower costs?

R possibilities

R Cloud - AT&T Labs

- Try It, Online or Locally - https://rcloud.social/tryit/index.html
  - support to different programming languages in the same code
  - code and analysis everything is public

Rstudio Server or connect???

- From "What is the difference between rstudio-connect and rstudio-pro" - https://community.rstudio.com/t/what-is-the-difference-between-rstudio-connect-and-rstudio-pro/43949

- RStudio Server Pro - Very similar to the Desktop IDE, but runs on a server.
  - it allows you to have more compute resources closer to your data,
  - it provides a uniform environment for teams that makes it easier to collaborate,
  - and it provides controls for administrators to monitor and scale work.

- RStudio Connect - Makes it easy to share R Markdown reports,
  - deploy shiny web applications, and APIs written in R.
  - GitHub stores your code statically, RStudio Connect knows how to run it, (accessibility to non R users... shiny applications)
  - run reports on a schedule and send emails with the results.

- R Studio Server on Google Cloud - https://towardsdatascience.com/r-studio-server-on-google-cloud-dd69b8bff80b

- Getting Started with RStudio Connect for GCP - https://support.rstudio.com/hc/en-us/articles/360033988434-Getting-Started-with-RStudio-Connect-for-GCP

Docker

- All starts with Docker;
- Primary focused on reproducibility? But also have your code running in any cloud service.

Container it! Rstudio with all you need in a docker container and deploy in a virtual machine... you don't need to install everytinhg again and again manually... and have your code "forever".

- You have to setup and install everything for the first time, start and stop the job whenever you want.
  - Even if you stop the VM if there is files in the storage will be charged by the service.

Everything is Docker!!!

containering R

- Docker + R - Rocker package - https://www.rocker-project.org/

- Rstudio...
  - support for Shinny apps

containering your code

- R.project + Github with Docker...
  - your favorite R packages
  - third party softwares... latte integrale for example

- how to add `C` library GMP, GNU for arithmetic precision? (part of latte distro) Via Rstudio Connect not working from its own terminal... use only docker file instead with R base???
  - latte distro - https://github.com/latte-int/latte-distro
  - https://gmplib.org/#DOWNLOAD (there is a R package that needs the distro installed)

open ssh in browser... packages not installed

```
sudo apt sudo apt update
sudo apt install build-essential
sudo apt-get install m4
```

upload latte distro .tar.gz (you can do this by clicks)... unpack and configure it!!!

- dockering R
  - Using R via Rocker A Brief Introduction to Docker for R - http://dirk.eddelbuettel.com/papers/chirug_nov2019_rocker.pdf
  - a good quick Docker introduction for R users - https://colinfay.me/docker-r-reproducibility/
  - example of calling additional libraries - Running your R script in Docker - https://www.r-bloggers.com/2019/02/running-your-r-script-in-docker/

- dockering R studio
  - a nice introduction and well organised setup - https://www.symbolix.com.au/blog-main/r-docker-hello
  - example setting additional libraries (but using rstudio server) - https://davetang.org/muse/2021/04/24/running-rstudio-server-with-docker/
  - sharing and Running R code using Docker - https://aboland.ie/Docker.html

https://code.markedmondson.me/r-at-scale-on-google-cloud-platform/

Example R and Google cloud

1. to create VM on Google cloud (after having an account)

a.. `VM instances > + Create` ... b. Open Cloud shell ... you can upload files/folders by clicking...

2. building and runing rstudio from your Dockerfile image

a. Build it

Type the following command into the command line. You must be in the same directory as your Dockerfile.

```
sudo docker build --rm --force-rm -t rstudio/my_simulation .
```

the –rm –force-rm just forces the container to delete itself once its scripts run or you log out. It just stops us filling up the server with lots of containers doing nothing.

`sudo docker image list` if you want to see your image added to the list.

b. Run it

'sudo docker run -d –rm -p 28787:8787 –name my_simulation rstudio/my_simulation

Done! Now open your webbrowser at `<yourhostip:28787>` . Username and password are both rstudio

Rstudio or R on google cloud???

- advantage in using rstudio images... maybe community is working more in containerisation... R Docker faster - "My experiment shows R Docker images will build much faster thanks to the new package manager from RStudio" from https://medium.com/@skyetetra/r-docker-faster-28e13a6d241d

- COSTS??? What would all this cost??

  - A VM with 1 vCPu n1-standard-1 (3.75GB RAM and 30GB SSD) with Rstudio from `Marketplace` costs around $5.5\times$ a VM from `Compute > Compute Engine > VM Instances > + Create VM instance` in my region;
  - So run your Dockerfile!!!

"the easy way" to get started quickly

(only if youre not confident yet to Docrerisation)

- Rstudio on google cloud Marketplace... https://support.rstudio.com/hc/en-us/articles/115010260627-Getting-Started-with-RStudio-Workbench-RStudio-Server-Pro-Standard-for-GCP

  - you can use it all interactively, but has to install everytime you setup a VM...
- installing tar.gz... via terminal

to unpack the bundle

```
tar -xvzf latte-integrale-1.7.3b.tar.gz
```

and to configure

```
./configure
make
```

Now all the files are in `/home/rstudio-user/latte-integrale-1.7.3b/dest`

install.packages("gmp") error... what is the best option? to submit jobs and use R distro base... or Rstudio facilities??? I couldnt install latte integrale distro interactively...

```
./configure
 make
```

Some self-tests can be run with

```
make check
```

And you can install (under '/usr/local' by default) with

```
make install
```

- Another quick way to launch Rstudio on GCP
    - Launch RStudio Server in the Google Cloud with two lines of R - https://code.markedmondson.me/launch-rstudio-server-google-cloud-in-two-lines-r/