Building and Managing Unified R Environments for Data Science and Software Development

Dariusz Ratman, PhD
Roche Global IT Solution Centre (RGITSC)

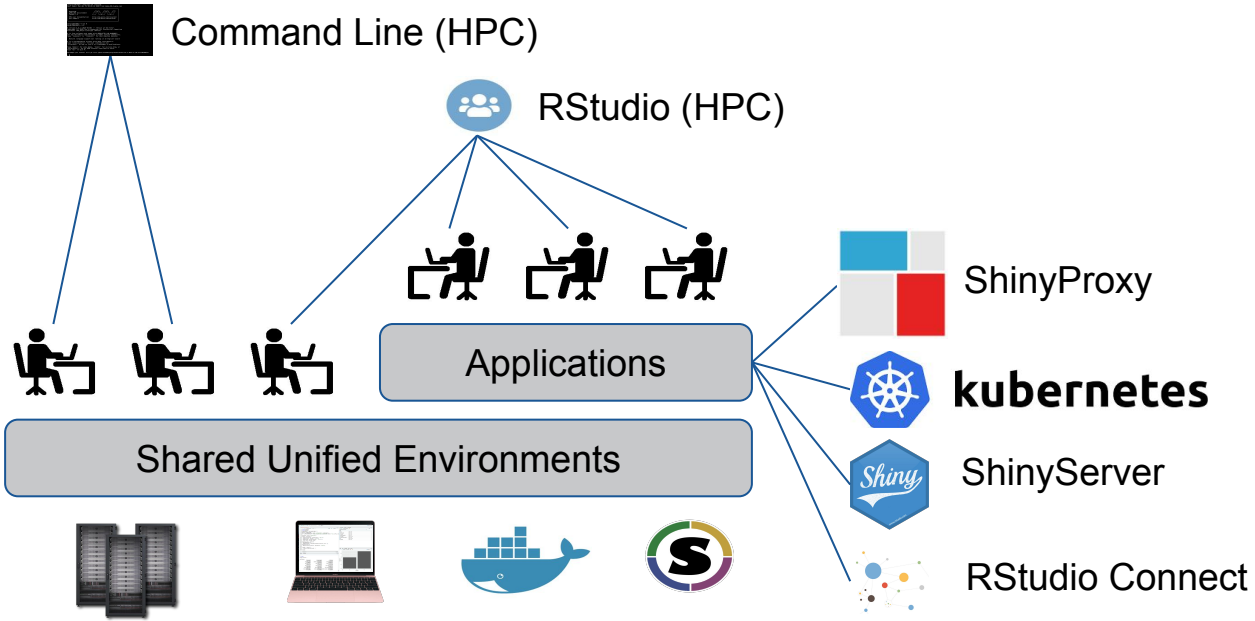
Computational Environment for Data Science & Research

Vision & Guiding Principles

*Provide **shared**, **unified** analytical computing environments that facilitate **reproducibility** and **comparability** while allowing for sustained **agility***

Computational Environment for Data Science & Research

R users & R ecosystem overview



R Users Community:

- Data Scientists / Analysts
- Software Engineers
- R Engineers
- Scientists

Computational Environment for Data Science & Research

The foundation

R Environment

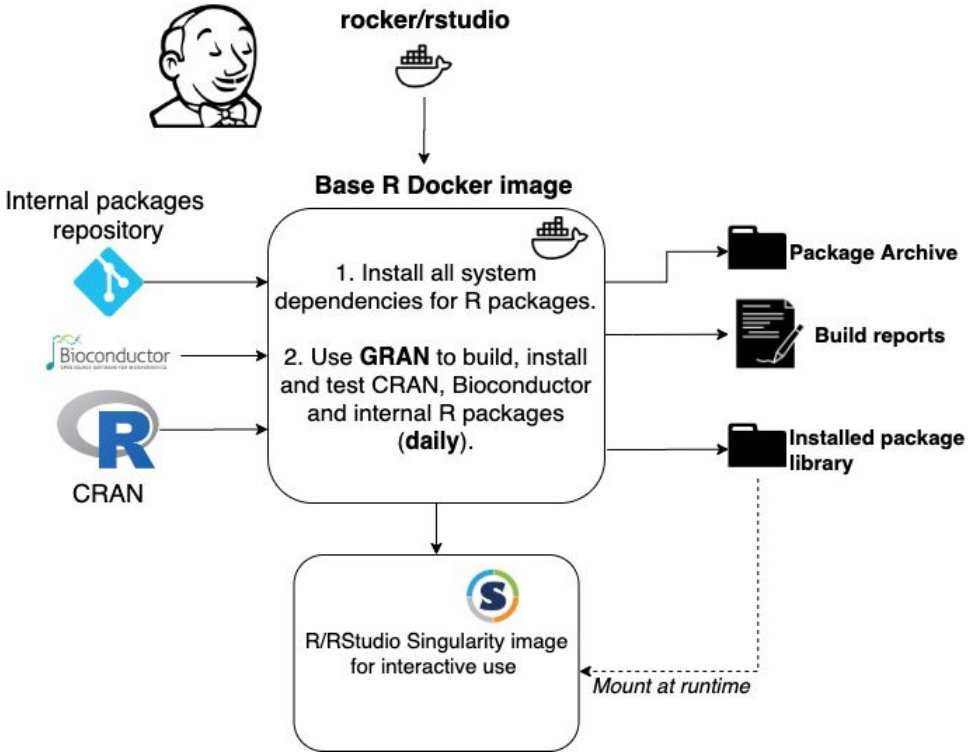
- Operating system + all dependencies
- R packages ~ **5000** available for Data Scientists / Analysts, including CRAN, Bioconductor, github and internally developed ones

Key Challenges

- All these packages need to “work” in a single environment
- **Reproducibility** requires environment versioning
- Environments should be **portable** so we can easily use them on different compute platforms and infrastructure

Computational Environment for Data Science & Research

The build system and release management

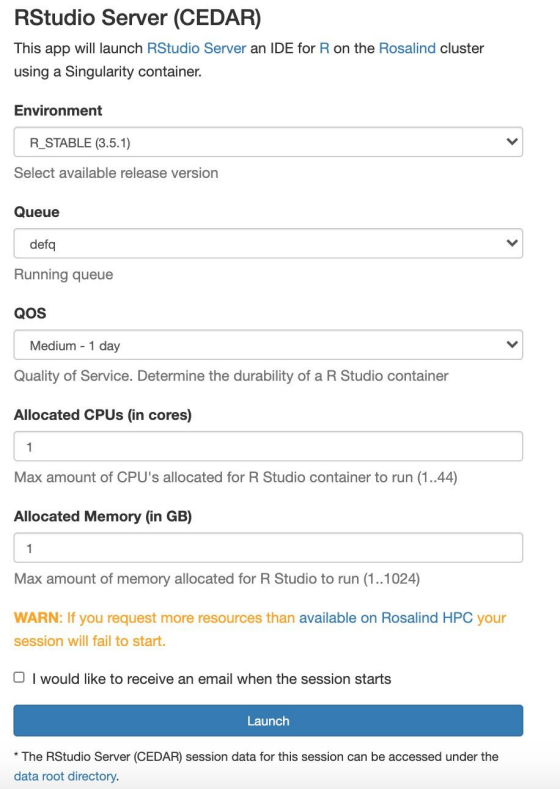


- Biannual **stable production** environment release following Bioconductor release cycle.
- Daily rebuild of the **TEST** and **DEVEL** environments.
- The environment CI based on containerized (Docker) GRAN instance(s) (still pre-production).

- Artifacts produced:
  - Package archive - exposed via web server.
  - RStudio Docker Image (base for development).
  - **Singularity image for use on HPC cluster.**
  - **Installed package library (external dir mounted to a container).**

R Environment for Data Scientists

User interfaces



- R command line interface on HPC cluster (Furlani module):

```
[ratmand@nl004 home/ratmand$ ml RP/singularity/R-3.6.1-bioc-3.10

Module RP/singularity/R-3.6.1-bioc-3.10/R-3.6.1-bioc-3.10 adds two commands:
R - to run interactive R terminal in Singularity container
Rscript <command> - to run R command in Singularity container and exit container
```

- Web app provided via **Open On Demand** platform to launch RStudio containers as batch jobs.
- In both cases user's container is spinned up from a common. (singularity) image and pre-installed R package library is mounted at runtime.
- Versioned releases with a corresponding **image** and **package library** enable reproducibility and can be relatively easy migrated to run on different infrastructure.

***CEDAR** - Compute Environment for Data Science And Research

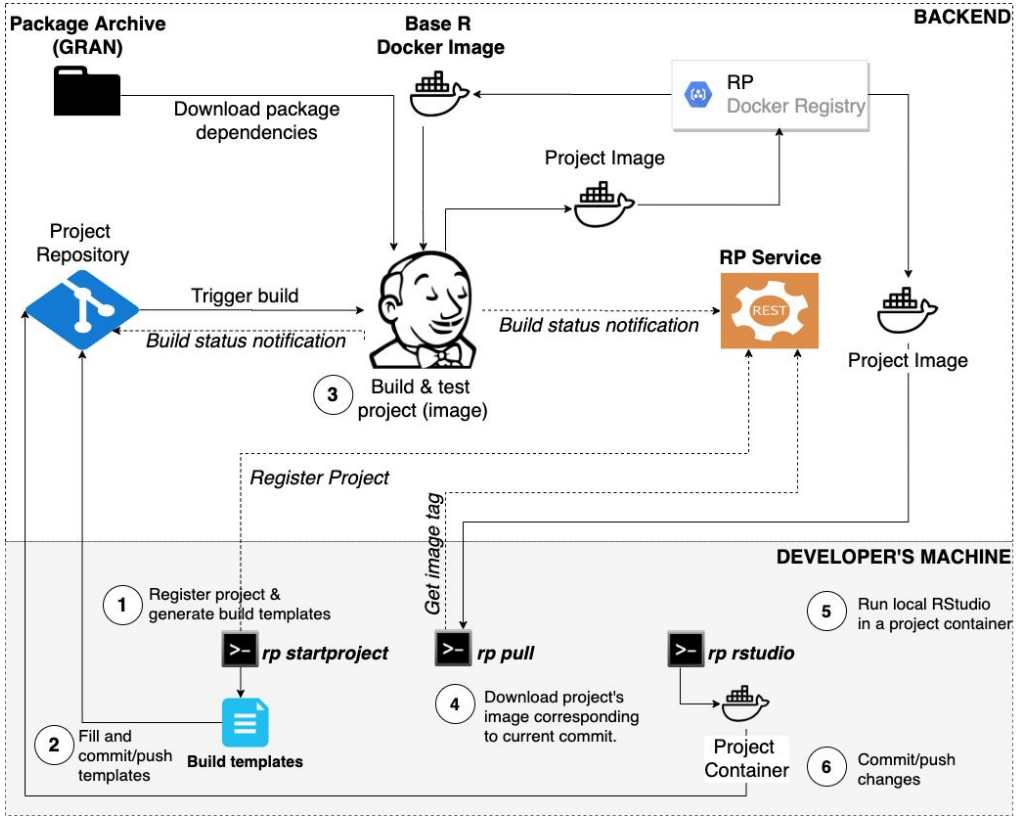
R Environment for Software Engineering workflows

Additional considerations

- Application development environments need to be **self-contained**, but should be consistent with the analytics environment with a corresponding R version (**comparability**).
- Development environments should be easy to create and share to enable collaborative work (**agility**).
- Larger development efforts require dedicated build and test infrastructure to enable **rapid feedback** (GRAN build system, which provides “nightly build” is too limiting for day-to-day development).

R Environment for Software Engineering workflows

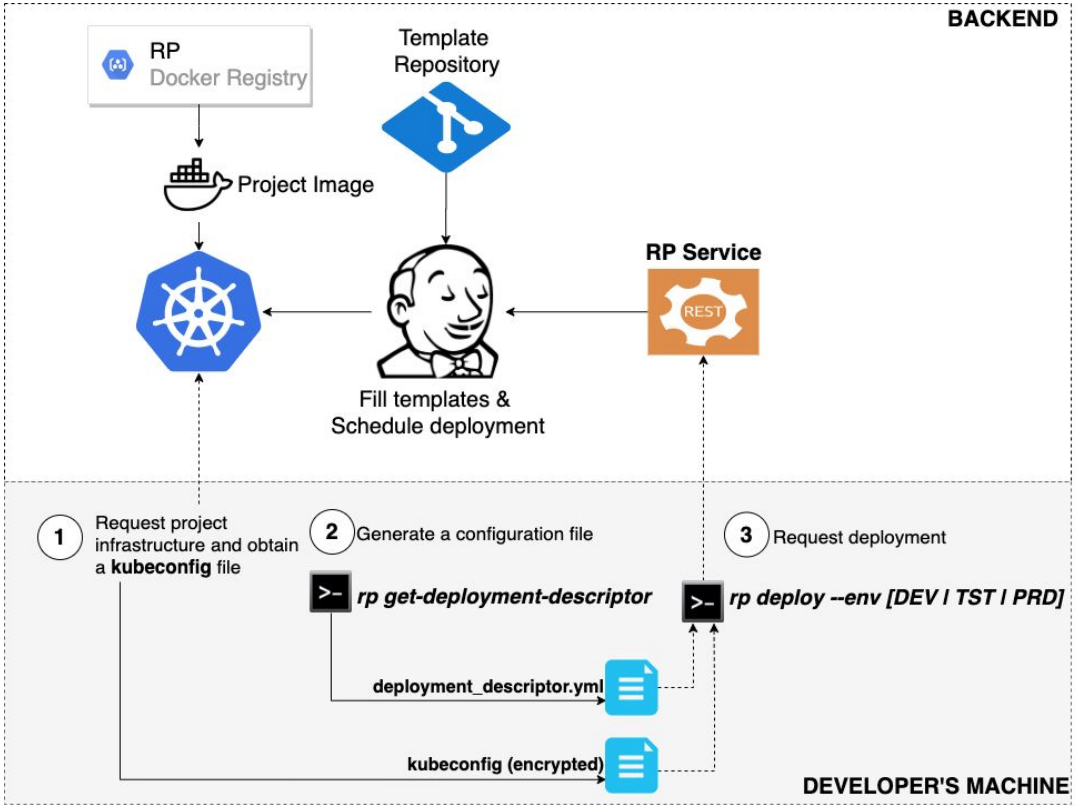
Development workflow & Tooling



- This workflow is well established among RGITSC **R engineering team** for development of R packages and Shiny apps with Docker containers.
- **rp CLI** - command line tool supporting the development workflow with git & Docker.
- **RPlatform (RP) Service**
  - Registers projects
  - Tracks which image tag corresponds to which git commit.
  - Provides a list of available base images
- **Base R Docker image** is configured to download packages from a corresponding **GRAN Package Archive** to enable environment consistency.

R Environment for Software Engineering workflows

Deployment workflow for Shiny apps



- Infrastructure provisioned on request (K8S) and *kubeconfig* file shared with the dev team.
- Application config (*deployment_descriptor.yml*) generated once using the **rp CLI** command.
- After configuration adjustment, **deployments** issued with a **single CLI command**.
- Shiny apps deployed behind **ShinyProxy** on **K8S** cluster.

Acknowledgments

Genentech Research & Early Development:

Michael Lawrence

Rena Yang

Gabe Becker (former employee)

Rafał Udziela

Roche Global IT Solution Centre

Artur Legan

Artur Starzyk

Bartłomiej Marcinkowski

Daniel Kierecki

Dawid Dacewicz

Gniewko Ostrowski

Henryk Popławski

Jakub Szukała

Kamil Foltynski

Konrad Dębski

Konrad Zieliński

Marcin Pączek

Patryk Szczęch

Paweł Piecuch

Paweł Przytuła

Tomasz Ziobrowski

Doing now what patients need next