

Managing an R infrastructure

Challenges and Experiences at Statistics Austria

Alexander Kowarik

Center Methods and Quality (CM)

Bernhard Meindl

Center Methods and Quality (CM)

Bukarest, 13.12.2023



The Setup

An overview about the R infrastructure

Server - Hardware and resources

- *mirrored Setup*: “Test-” and “Production Server”
 - fully virtualized and snapshotted
 - Test-Server: 8 Cores / 128GB RAM; Prod-Server: 18 Cores / 378GB RAM
- *additional vServers* for computer-intensive projects
- *Build-Server* (running *Jenkins*) to check/build internal R-packages
- *Jfrog Artifactory*: Mirror/Proxy for external and internal R Packages
- internal *BitBucket* Server for Git

Note

Hardware-maintenance and Software updates are typically done by IT

Software - What do we use?

- **R**: every odd **R** version (e.g, **4.2.3**, **4.3.1**) is installed
- **Posit Workbench** + **Connect**: are updated timely;
 - a few days between update on Test- and Production server
 - Updates are announced via Email in advance
- **Server OS**:
 - a stable (Ubuntu Server LTS, currently **20.4.6**) system is chosen
 - security updates are installed every few month



Tip

Keeping the software and tool up-to-date is crucial

A photograph of a modern office building interior, featuring multiple levels with glass railings and indoor plants. The image is overlaid with a semi-transparent blue filter. On the right side, a window with white blinds is visible, showing a view of another building.

R-Users

Add, manage and support Users

Adding Users (1)

- **Necessary steps** until a user gets access to the R-infrastructure
 - Authorization via internal “*Identity & Access Management*” Tool
 - a single permission required to gain access to **Workbench**, **BitBucket**, **Jira** and **Confluence** (Wiki)
 - extra permission required to get access to **Posit Connect**
 - Ticket (**JIRA**) must be created for IT
 - local account (with **\$HOME**) needs to be created at the relevant **R** Server(s)
 - Users must be notified that they can start using the **R** infrastructure

Adding Users (2)

- we try to **facilitate** and **automate** the process as much as possible
- **Authorization** via IAM must be granted manually, but ...
 - **Jira-Ticket**: required for IT is generated automatically
 - How? a scheduled Rmd-Job (running on **Connect**) checks for updated permissions in relevant IAM-groups
 - **Notification**: Onboarding-Email with general Information (Links to Servers, Wiki with **R** content, Jira, ...) is automatically sent
 - How? a scheduled Rmd-Job (running on **Connect**) checks and keeps track of new User-Accounts on the **R**-Servers
- **Future-Tasks**
 - How to deal with accounts/data from users that no longer work at STAT?

Supporting Users (1)

Tip

- Having good accessible documentation reduces support-time
 - It is helpful to be able to refer to existing documentation
-
- **Confluence** (Wiki): all R Users have automatically access
 - FAQs are maintained
 - Important questions (e.g how to restrict access to shiny apps) are documented
 - **Jira**: all R Users can create and view issues in topic **RSUPP** (R-Support)
 - we encourage colleagues to ask questions via issues (and not via Mail/Phone)
 - any colleagues can provide helpful answers to improve response times

Supporting Users (2)



Tip

The easier it is for users to start, the more likely it is that they will use R

- **Utility-packages:**
 - make life easier for users
 - also help to maintain standards and avoid re-implementation of core-tasks
- **Pipelines:** a set of pipelines is maintained and developed
 - R (package) → Jenkins → Artifactory
 - R (shiny/plumber,rmd) → Jenkins → Connect

Supporting Users (3)

- **Subset** of internal packages:
 - **authSTAT**: securely store secrets (e.g DB creds)
 - **dataSTAT**: harmonize access to databases
 - **mountSTAT**: allow access to windows-based file shares
 - **useSTAT**: create projects, interact with Jenkins
 - **apiSTAT**: interact with various APIs (**Jira**, **Connect**, **Bitbucket**, **Jenkins**)
 - **slideSTAT**: create slides in corporate design
 - **rinstSTAT**: install/query **R** versions, create issues, ...
 - **sampSTAT**:` utilities to perform sampling-related tasks
- → updated nightly for latest two **R** versions



Workload

... How to manage / monitor?

Manage Workload (1)



Being able to monitor the workload helps to prevent service stoppage

- Possible **reasons for failure?**
 - full filesystems (`$HOME`, `/tmp`)
 - overloaded CPUs
 - Out-of-Memory Killer

Manage Workload (2)

- **Monitoring:**
 - via automated, periodic Cron-Jobs that (in case) send mails
 - Scripts query a Plumber-API that returns current information / stats
 - free RAM
 - CPU usage
 - available space in important file-systems
 - overall open sessions per user

Manage Workload (3)

- **Allocation** of resources
 - both **Workbench** and **Connect** have internal features to limit resources
 - related to users (e.g number of concurrent sessions, RAM, ...)
 - related to processes (relevant for **Connect**)
 - **Quotas**: limits related to allowed space on \$HOME per user have been implemented
 - **CPU-priorities**: in case users choose to explicitly use parallelization, they need to **(re)nice** their process (documented in Confluence)



Other aspects

... What else do we have or work on?

What more do we have / work on?

- **Admin-App:** a Shiny-App deployed on **Connect** that allows **R-Admins** to create standardized Jira-Tasks for IT staff
 - Force Password reset
 - Update storage groups (how many RAM a user may allocate)
- Simple Object Storage (**sosSTAT**) with a database backend (similar to **pins-**package) that will allow versioned storage of files



The Future

... where do we want to go?

Future tasks / ideas / goals

- Identify and remove possible single points of failure
- Split servers for **Connect** / **Workbench**
- Possibly change licensing model for Posit products
- Explore possibilities (and problems) of containerization
- Improve User-Participation (possibly organizing an internal “*conference*”)

Note

Thank you for your attention!

Please address queries to

Alexander Kowarik

Center Methods and Quality (CM)

✉ Alexander.Kowarik@statistik.gv.at

Bernhard Meindl

Center Methods and Quality (CM)

✉ Bernhard.Meindl@statistik.gv.at

STATISTICS AUSTRIA

Guglgasse 13, 1110 Vienna

What is the awesome list?

Curated list of software for official statistics



www.awesomeofficialstatistics.org

GitHub navigation bar with search bar and links: Pulls, Issues, Marketplace, Explore

Repository header: SNStatComp / awesome-official-statistics-software. Interaction buttons: Unwatch (30), Unstar (161), Fork (41)

Social interactions

Repository navigation: Code, Issues, Pull requests (1), Actions, Projects, Wiki, Security, Insights

Repository controls: master branch, Go to file, Add file, Code button

Awesome official statistics software

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

Criteria

An item on this list is awesome because

1. it is free, open source, and available for download and
2. it is confirmed to be used in the production of official statistics by at least one institute or it provides access to official statistics publications.

We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. [Contributions](#) are welcome.

License



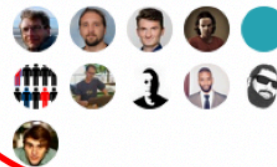
This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

Open license

An awesome list of statistical software for creating and accessing official statistics

official-statistics gsbpm

Contributors 15



+ 4 contributors

Working together

Contributions

Awesome contributions are welcome, here are ways to do it:

- The GitHub way: send us a [pull request](#) to add directly to this list.
- Add an item to the [issue tracker](#) issue tracker. (you need a GH account)
- Send an e-mail to [mark dot vanderloo at gmail dot com](mailto:mark.vanderloo@gmail.com) or [olav dot tenberch at gmail dot com](mailto:olav.tenberch@gmail.com) or tweet [@markvdloo](https://twitter.com/markvdloo)