DESTATIS
Statistisches Bundesamt

# Use of R and SAS at the German Federal Statistical Office

Dr. Andreas Jahn

Nov. 25th, 2024

# SAS and R in comparison

5 SAS installations, SAS 9.4. are currently in use:

- SAS is used for the national census project for data evaluation
- About 1.000 users (stable), out of 2.500 total employees at destatis
- About 100.000 SAS projects
- Guidelines for the SAS programming are in use

- Update of the current server beginning of 2025
- R was used in the national census project for data extrapolation
- About 300 user, growth expected up to 450
- Guidelines for R-code development is about to be published

# R-Server, technical details

**Load Balancing Cluster** (Bare Metal) with 10 x (2 x NVIDIA RTX-A6000 P 48GB VRAM + 2 x 22 Cores CPU + 2TB RAM)

**Installed Software**: Posit Workbench Advanced (RStudio, Jupyterlab, VSCode), R, Python

**20 TB NFS Storage**

**Gitlab** Server (Community Edition) for CI/CD

Posit Package Manager for **Package Management** (Python (PyPI) & R (CRAN, Bioconductor))

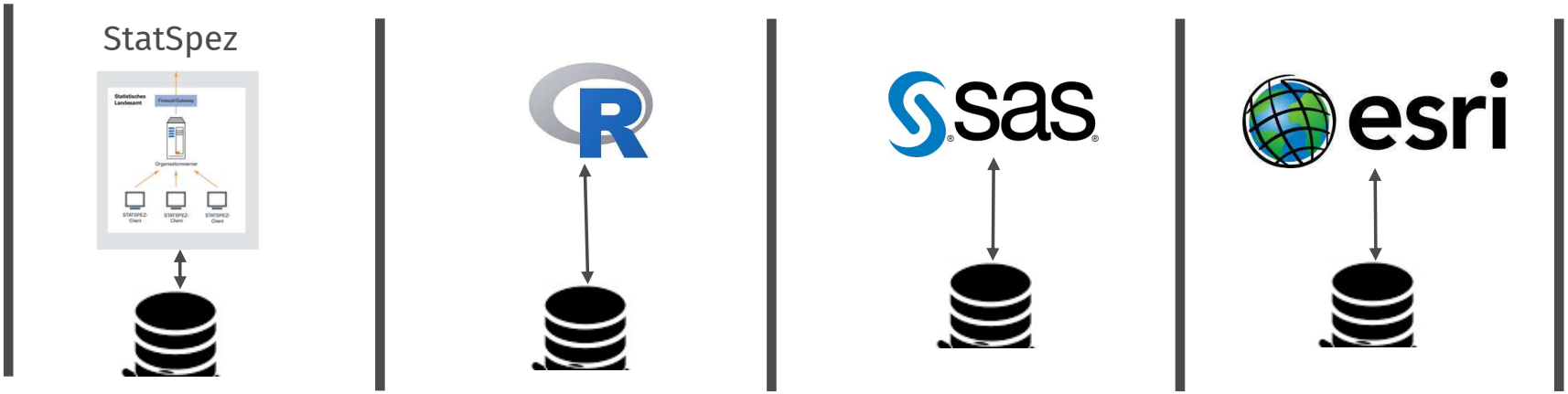
**OS**: Red Hat Enterprise Linux 8.10

# Current initiatives

- **Run R- and Python Code in a SAS Viya environment:**
  - Basic R-Code can be used in a SAS environment, Python is better supported
  - But:
    - Package management is complicated
    - SAS Viya does not provide a proper development bench for R code

- **Use Large Language Models (LLMs) for automated translation of SAS and R Code**
  - Automated translation from and to SAS Code is possible
  - For complex programs automated translation is not advised
  - LLM are great for analyzing old (undocumented) code

# Current state

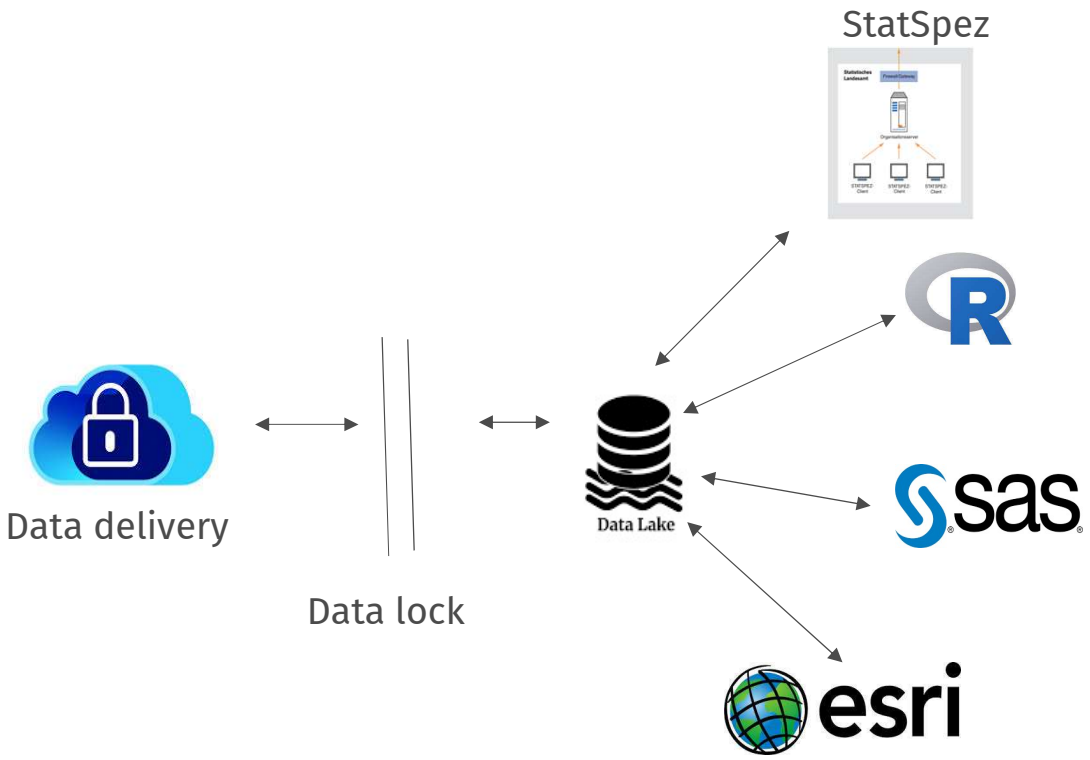



- Servers used for analyzing data are separated.
- Data is stored on the corresponding servers, exchange of data cannot be done directly, only via file-transfer and data locks.
- Identity and user management are not harmonized.

# Target scenario

Analysis servers have access to a common data lake.

Servers have access to the same data for manipulating purposes.



# Conclusion

- Usage of **R** is becoming increasingly **popular**
- A **huge legacy of SAS** programms exists, a complete migration at short notice is not possible, migration will be costly
- **Different tools** for data analysis exist in parallel and will continue to do so
- More emphasis must be put on **a common data-management infrastructure**, making it easy to replace tools while proctecting the data