OSSDIP: Open Source Secure Data Infrastructure and Processes Platform that Supports Data Visiting

Martin Weise [1] Tobias Grantner [1] Lukas Mahler [1] Andreas Rauber [1]

Project Website
ossdip.at

Problem Description

In an increasing number of settings, both researchers in academia as well as stakeholders in industry need to **safeguard access** to highly sensitive data.

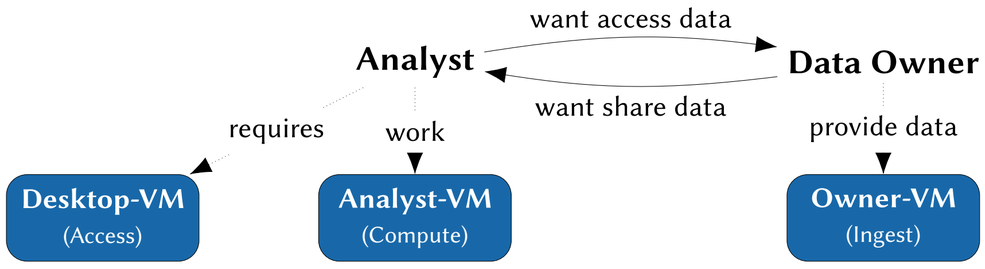

Figure 1: Social architecture (highlight)

While data sharing is being proclaimed as the future in open science, many settings **do not allow** for such approaches due to e.g. confidentiality concerns.

Methodology

Based on the UK HDRA Trusted Research Environments (TRE) definition and experience of operating DEXHELPP for almost ten years. Provide **highly controlled** and **monitored** data visiting services, without disseminating an actual copy:

- Components of *data anonymization* and *fingerprinting*
- Extensive *logging* and *monitoring*
- Defined *processes* and contractual frameworks

Secure Data Infrastructure

The overall concept is centered around the principle of **never providing access** to the **data node** where all data is being held. For each individual analysis request:

- Specific subset of the data required is extracted from the data node, and
- Copied onto a dedicated Analyst-VM, together with the tools required to perform the analysis

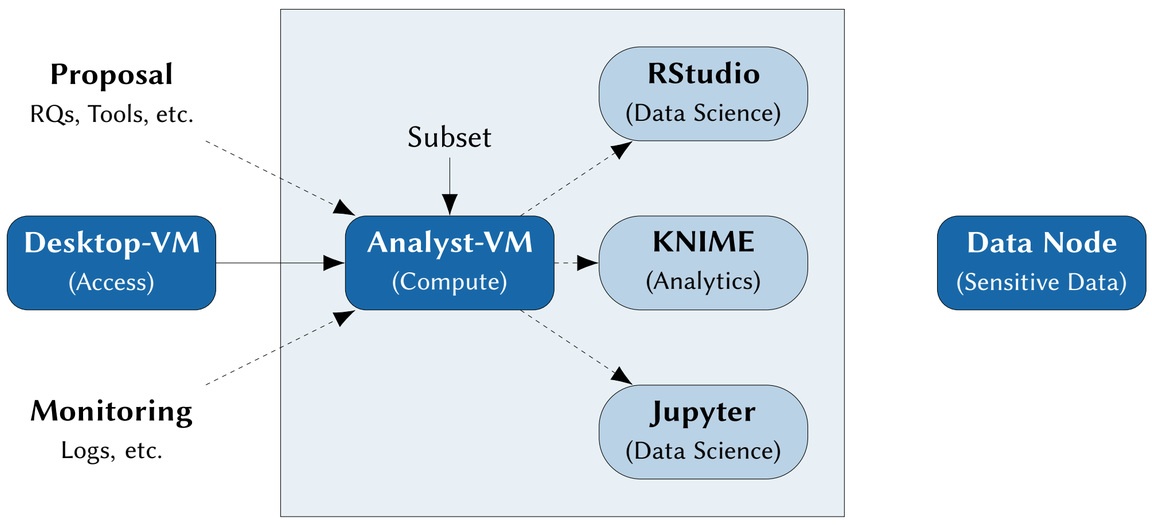

Figure 2: Data access workflow (highlight)

Access to this Analyst-VM is granted to the analyst working on the task at hand – however, **never directly**, but only via a dedicated Remote Desktop-VM to introduce a **media break** and avoid any data flowing off via e.g. a tunnel.

Research Activities via Data Visiting

A **dedicated** Remote Desktop-VM is created to provide the sole access to the Analyst-VM. The Analyst then can analyze the data as long as the time-out is not reached.

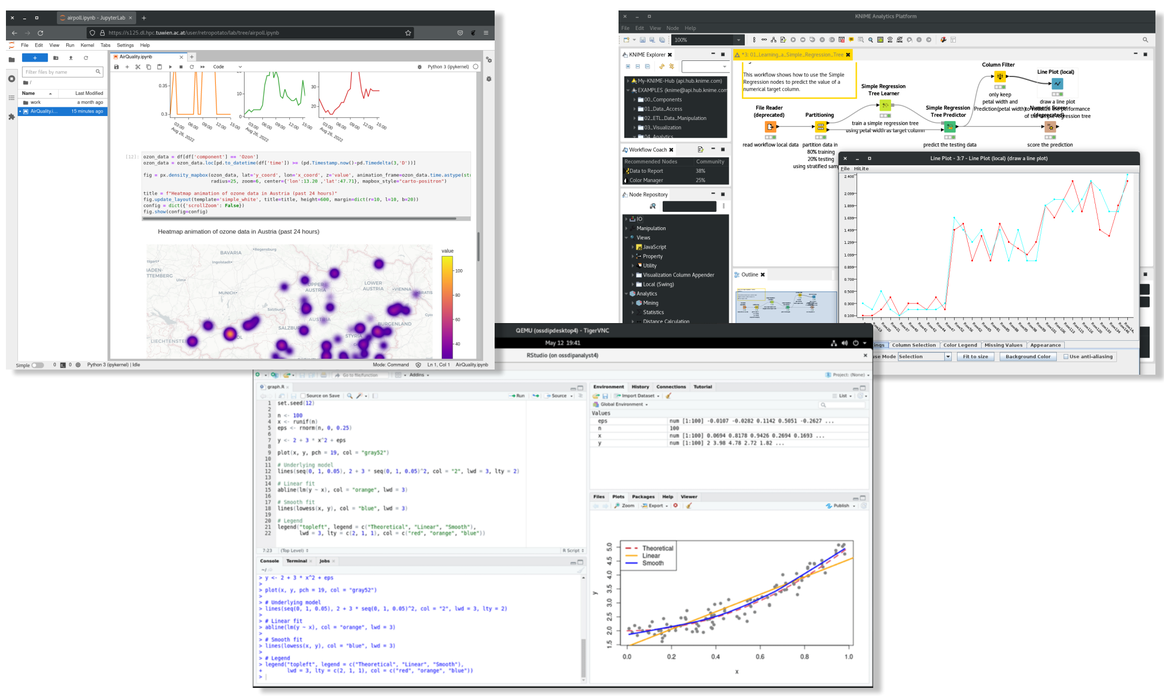

Figure 3: Working with the raw data in Jupyter, RStudio and KNIME

Our reference implementation supports **open-source** tools by default:

- *Data science*: Jupyter Notebooks, RStudio, KNIME
- *Text processing*: Libreoffice, LaTeX
- *Programming*: Python, Java, R

It can be extended to support commercial software as well in the configuration files

Conclusions

The current state of the secure data infrastructure allows a Data Owner to **invite experts** (e.g. Analyst) to visit sensitive data on a **trusted meeting point** (Analyst-VM). OSSDIP requires at least three (optimally nine) physical machines on trusted hardware.

Contact

Technical lead Projektass. Dipl.-Ing. **Martin Weise** looks forward for your e-mail: martin.weise@tuwien.ac.at

References

[1] Project Website. https://ossdip.at/.

[2] M. Weise, F. Kovacevic, N. Popper, and A. Rauber. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. *Data Science Journal*, 21:4, 2022. doi:10.5334/dsj-2022-004.

[3] M. Weise and A. Rauber. A Data-Visiting Infrastructure for Providing Access to Preserved Databases that Cannot be Shared or Made Publicly Accessible. In *Proceedings of the 17th International Conference on Digital Preservation*, 2021. doi:10.17605/OSF.IO/VKN4R.

[1]Vienna University of Technology