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.

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

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.

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

