A data-visiting infrastructure for providing access to preserved databases that cannot be shared or made publicly accessible

Martin Weise
TU Wien, Austria
martin.weise@tuwien.ac.at
0000-0003-4216-302X

Andreas Rauber
TU Wien, Austria
andreas.rauber@tuwien.ac.at
0000-0002-9272-6225

Databases preserved in archives contain highly valuable information that frequently cannot be made freely accessible for analyses via standard data portals, be it due to legal, commercial or ethical issues. We present Open Source Secure Data Infrastructure and Processes (OSSDIP), the reference implementation of a high-security data visiting infrastructure initially conceived as a safe-compute environment for medical data. It provides highly controlled and monitored data visiting services while ensuring to the largest degree possible that data cannot be extracted from the infrastructure. This may offer archives a viable alternative for providing restricted access to sensitive data in a more flexible manner.

Keywords – Secure Data Infrastructure, Data Visiting, Data Dissemination

Conference Topic – Enhancing the Collaboration

I. Introduction

Providing access to archived databases is an ongoing challenge for many organizations where simply providing access to archived data or handing over the data is not feasible. This may be due to privacy reasons or commercial sensitivity of data. Yet, they still may want to make the data accessible to third parties for research purposes or to assist with specific analytical tasks. Commonly, this is resolved by visiting the organization onsite and work with the data within the organization's premises, or by conducting a range of non-disclosure processes. We propose a data access set-up that allows an archive to provide access to the data without disseminating an actual copy, but by inviting third parties to visit the data electronically within a controlled environment. To enable this we borrow concepts and infrastructure set-ups from domains where such access provisioning to highly sensitive data is a common setting: among many other domains, medical data repositories frequently have to operate in such settings, implementing principles such as the five saves model [1]. Adapting these to the needs of archival institutions may allow them to make their data accessible in limited, strictly monitored environments, allowing users to work with the data while ensuring that the archive remain in complete control of the data and specifically that these records are not being exfiltrated.

Meeting privacy requirements when passing data to external service providers as part of a digital preservation solution may be met e.g. by applying encryption on the data [2]. Yet, this is not easily feasible for providing access to sensitive database records for analytical purposes – with the exception of homomorphic encryption, which, however, severely limits the analytical activities that can be applied on the data.

We present OSSDIP, an open source reference implementation facilitating access to sensitive data in monitored data visiting settings. Originally developed for providing analysts access to highly sensitive medical data it may serve similar purposes in other settings where archives want to provide access to sensitive data.

II. Infrastructure

Our secure data infrastructure stores the sensitive data in a dedicated, shielded server, instantiated as a Virtual Machine (Data-VM) in our reference implementation, that has a strict firewall barrier around it. Only process-approved connections to selected VMs, as well as for maintenance and Monitoring-VM are allowed to pass through this. Yellow components in Figure 1 are secured by at least two security layers (e.g. connections passing the VPN-VM need also SSH or are encrypted a second time via the Remote Desktop encryption as described further below). Everything right of the VPN-VM including itself can be virtualized on a single Virtualization Host for demonstration purposes, but will usually be deployed on several physical machines for scalability and added security reasons.

The overall concept, in a nutshell, is centered around the principle of never providing direct access to the server where the sensitive data is being held

17th International Conference on Digital Preservation
iPRES 2021, Beijing, China.

DOI: 10.1145/nnnnnnn.nnnnnnn

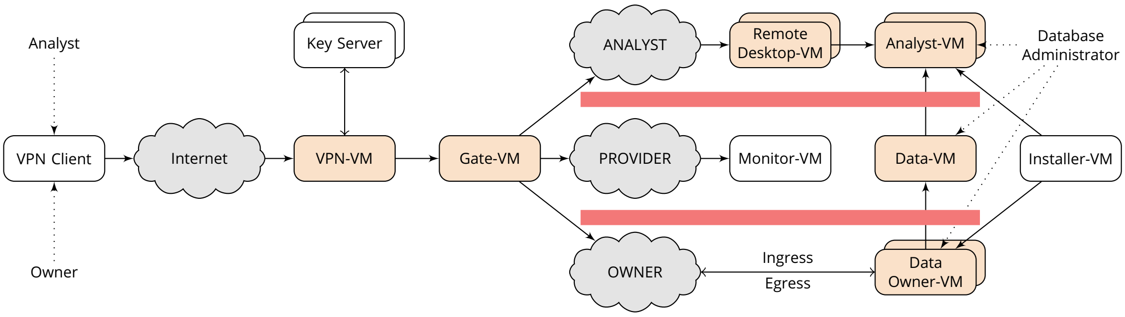

Figure 1: Overall system architecture

(i.e. the Data-VM). Instead, for each individual access request, the data required is extracted from the Data-VM and copied onto a dedicated analysis virtual machine (Analyst-VM) together with the tools required to perform the analysis. Access to this Analyst-VM is granted to the individual Analyst working on the task at hand – however, never directly, but only via a dedicated Remote Desktop-VM to introduce a media break which prevents any data flowing off through e.g. a tunnel while adding a secondary encryption layer on top of the VPN encryption. Furthermore, the video connection can be fully monitored and recorded to identify attempts on data theft. An Analyst establishes a remote desktop connection to a dedicated VM from where he or she has the sole possibility of connecting via secure shell (SSH) to the corresponding Analyst-VM, which, in turn, holds only a copy of the required and permitted subset of data (possibly finger-printed, aggregated, or pseudonymized).

This may seem like a major resource overkill approach in contrast to a e.g. PostgreSQL view on a data set, but ensures that even in presence of software bugs, the Analyst will never see more than the Data Owner approved. Additionally, our solution enables direct data manipulation on a copy without the need of creating new tables, procedures, etc. which is prone to human error. Export of any result files (trained models, figures, charts) is again possible only via a dedicated Data Owner-VM and explicit approval of the Data Owner. These VMs are being initialized with 0-ed, encrypted storage and destroyed after a specific transfer or analysis task has been completed.

III. Roles and Processes

Complementing the technical enforcement processes regulate data access requests and the transfer of credentials upon approval. Legally binding terms of use to allow the processing of personal information, non-disclosure agreements, extensive monitoring, or data access agreements complement the technical set-up with the following key roles:

The *Analyst* identifies what part of the sensitive data held by an archive (identified via metadata published by the archive) is required to answer a specific research question. With the permission of the Data Owner, the Analyst is able to work with that subset of data in an isolated VM. The *Data Owner*, i.e. the archive holding the preserved databases, has a strong interest in providing access to the data but wants to retain control of the data and specifically reduce the risk of data leakage to an acceptable level. The *System Administrator* maintains the secure data infrastructure environment, except for the Data-VM. This role also manages the platform where the infrastructure is hosted. The *Database Administrator* is responsible for maintaining the Data-VM and the PostgreSQL database running on it.

IV. Conclusions & Future Work

We have presented an architecture as well as a reference implementation which allows archives to make sensitive data from their holdings accessible to analysts for specific tasks while maintaining full control of how their data is being used and preventing data leakage. The software package as well as documentation is available on Gitlab[1]. The set-up runs largely automated via Ansible-playbooks. Future work will include the addition of passive security measures such as aiming to integrate real-time on-screen watermarking [3], [4] (for the Remote Desktop-VM) to detect data theft and leakage via screen capturing as well as evaluating practicability with organizations and real-archival data.

References

[1] United Kingdom Health Data Research Alliance, “Trusted Research Environments,” Version 2.0, 2020.

[2] A. Berdini, “Digital preservationina high security environment: Student records, encryption, and preservation,” in *Proceedings of the 16th International Conference on Digital Preservation (iPRES2016)*, Amsterdam, The Netherlands, 2019.

[3] M. Piec and A. Rauber, “Real-time screen watermarking using overlaying layer,” in *2014 Ninth International Conference on Availability, Reliability and Security*, 2014, pp. 561–570.

[4] D. Gugelmann, D. Sommer, V. Lenders, M. Happe, and L. Vanbever, “Screen watermarking for data theft investigation and attribution,” in *2018 10th International Conference on Cyber Conflict (CyCon)*, 2018, pp. 391–408.

[1] `https://gitlab.tuwien.ac.at/martin.weise/ossdip`