Secure Processing of Sensitive Data on shared HPC systems

Scheerman, M.[*], Voort, L.[*], Zarrabi, N.[*]

[*]SURFsara, SURF Netherlands

Summary

In this work we present a novel method for creating secure computing environments on traditional multi-tenant high-performance computing clusters. Typically, current HPC clusters operate in a shared and batch mode, which can be incompatible with the security requirements set by data providers for processing their highly sensitive data. We propose a solution using hardware and network virtualization, which runs on an existing HPC cluster, and at the same time, meets strict security requirements. We show how this setup was used in two real-world cases. The solution can be used generally for processing sensitive data.

Background

High-performance computing (HPC) clusters have proven themselves as an effective solution for processing large volumes of data. With the increasing availability of confidential data, and the growing awareness of data security and privacy, there is a high demand for HPC facilities with a tightly managed security level.

One large source of confidential data is the Dutch government. The Central Bureau for Statistics (CBS) in The Netherlands acts as a data steward for the Dutch government-owned confidential data. The Dutch government has strict laws about processing and disclosure of its data. CBS ensures compliance with the law, by strictly regulating access to data, disclosure of research results, and having data processor agreements signed with users of the data. In practice, this means that only authorized researchers can access the confidential data in a tightly controlled environment. They can only take the results of their analyses out of the controlled environment after inspection and approval by a CBS data manager.

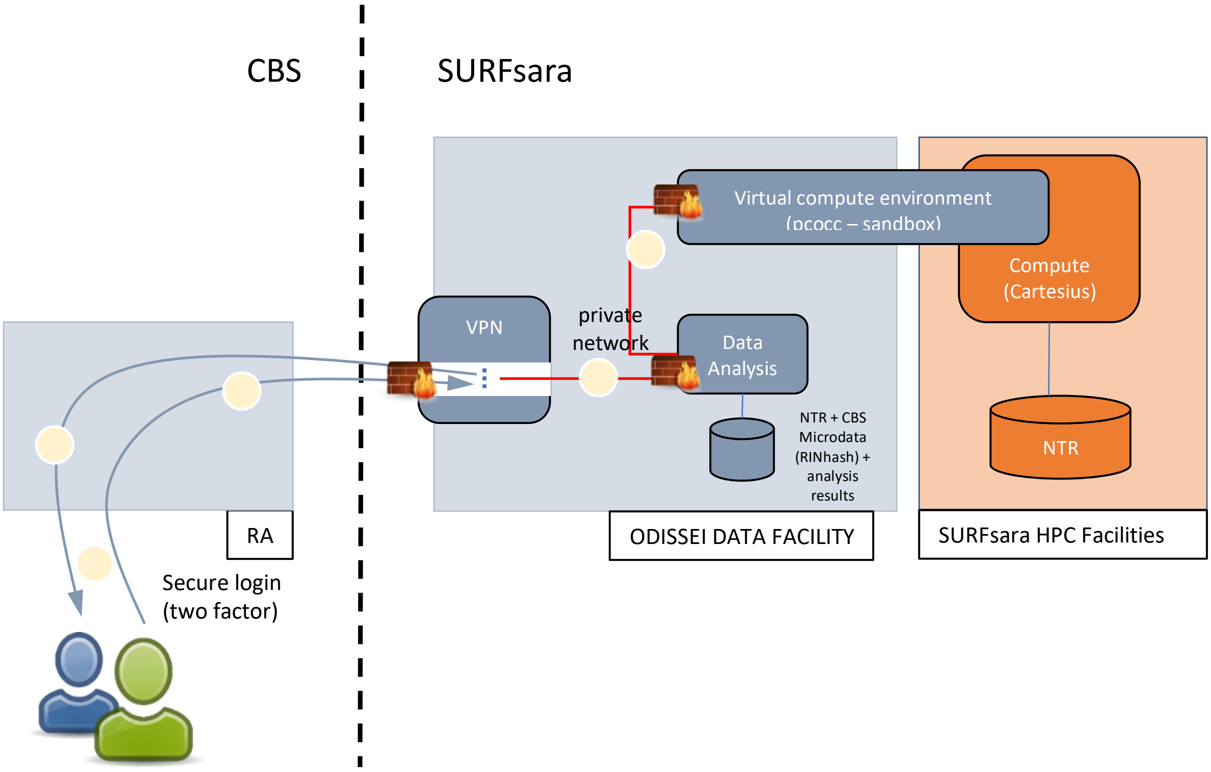
Two research projects were proposed by CBS, in which highly sensitive government-owned data (CBS Microdata) was to be used. One project was in the field of social geography, and the other project was in the field of healthcare, where the genetic dataset registered in the Netherlands Twin Registry (NTR) needed to be linked to the Microdata.

SURFsara is the Dutch supercomputing centre, and maintains and operates different compute clusters for academic users. Cartesius is SURFsara's supercomputer with 2000 multi-core compute nodes, connected with Infiniband for high-speed, low-latency communication and storage access. Users have a wide selection of pre-installed and maintained software. Cartesius offers multiple Terabytes of storage for projects, making data-intensive research possible. For academic users, the security level offered by typical clusters suffices. However, the two projects in which large volumes of government-owned data were needed, required additional, more strict security measures. SURFsara was asked to develop a secure high-performance compute facility, where users could use the compute resources and software offered on the Cartesius supercomputer; but with a level of security meeting the requirements as set by the data owner.

Work carried out

SURFsara, in collaboration with ODISSEI (http://www.odissei-data.nl/) and CBS (https://www.cbs.nl/), has built a secure platform for linking, analysing and processing sensitive data on HPC. The platform is based on customisable virtualized private clusters that are deployed on the existing supercomputer Cartesius. The virtual clusters are automatically provisioned using PCOCC (Private Cloud on a Compute Cluster), developed by CEA (https://github.com/cea-hpc/pcocc). The PCOCC technology allows users of an HPC cluster to host their own cluster of virtual machines on existing compute nodes. PCOCC itself uses Linux host virtualisation technology, OpenvSwitch for network virtualisation, and integrates with the SLURM job scheduler for deployment. For the two presented use cases, the network was configured such that access was only possible through a VPN connected to the data owner, and that data could not leave the system except via the data owner. After the data processing tasks are finished the virtual cluster is destroyed and the data is cleared.

In one use case, SURFsara acted as a *Trusted Third Party*. In this case, SURFsara manipulated data from one source such, that sensitive information could not be inferred from combining datasets, but that at the same time, the data could still be used for the research. Using test data, we showed that results were not changed by the TTP manipulation process.



Results

The resulting setup met both functional and security requirements and made it possible to store and process and analyse highly sensitive data on HPC. The two use cases were carried out successfully with respect to heavy computations; in the first use case, a speed-up of about 40 was achieved, reducing the compute time from about one year to about one week. In the second use case, combining the datasets was made possible by this setup, and the heavy computations and analyses were carried out successfully.

Conclusions and Future work

In this work we have addressed the problem of processing sensitive data on shared HPC systems. This work, to the best of our knowledge, is the first attempt to be able to store and process sensitive data on shared HPC systems without modifying the existing infrastructure. Further we are looking into secure storages options that can be connected to the platform for long term storage of sensitive data. We are also planning to perform a security audit test on the final design of the platform.

References

1) Europe PMC Funders Group. 2014. “Biological Insights From 108 Schizophrenia-Associated Genetic Loci.” *Nature* 511(7510):421–27. Retrieved July 5, 2018 (https://www.nature.com/articles/nature13595).
2) Petrović, Ana, Maarten van Ham, and David Manley. 2018. “Multiscale Measures of Population: Within- and between-City Variation in Exposure to the Sociospatial

Context.” *Annals of the American Association of Geographers* 108(4):1057–74. Retrieved October 9, 2018 (https://www.tandfonline.com/doi/full/10.1080/24694452.2017.1411245)