Secure Processing of Sensitive Data on shared HPC systems

Narges Zarrabi

CompBioMed Conference– 27th September 2019

SURF

Challenges - Sensitive Data

- Sensitive data
  - Must be protected against unwanted disclosure
  - Access should be safe guarded
  - Protection required for legal or ethical reasons
  - Personal privacy

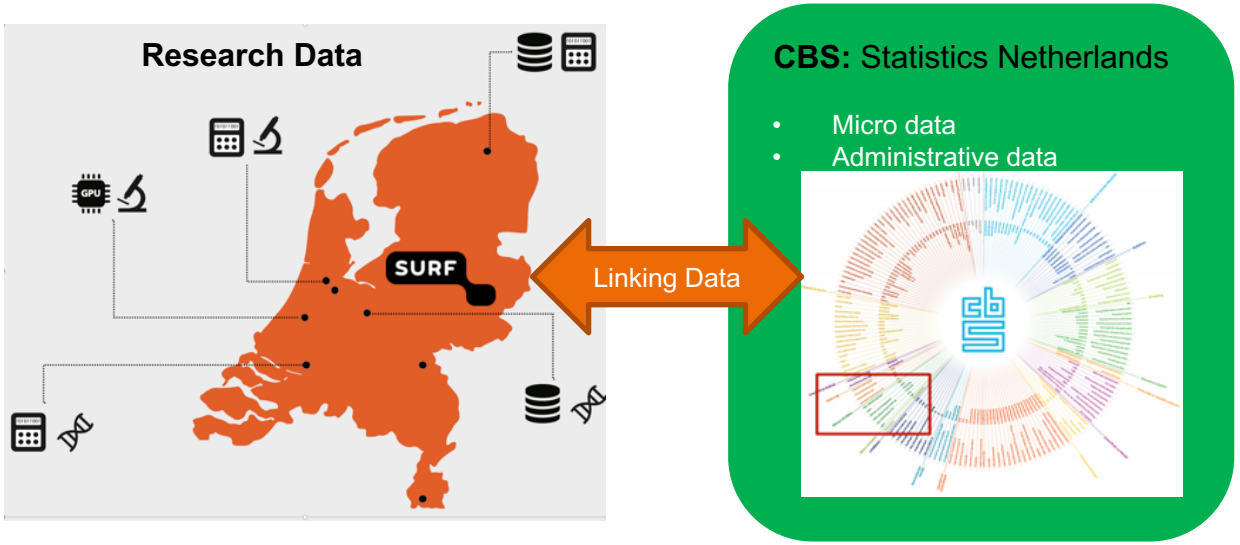
- Increasing availability of confidential and sensitive data, and the growing awareness of data security and privacy, urges the demand for HPC facilities with a tightly managed security level.

- Current HPC clusters often operate in a shared and batch mode, which poses challenges for processing sensitive data

ODISSEI Community

- ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations)

- Research
  - Economics
  - Political science
  - Human geography
  - Sociology
  - Psychology
  - Demography
  - Law



- Data
  - Socio-economic, demographic, health, crime, income and wealth variables of the complete population of the Netherlands (17 million)
- Challanges:
  - Processing sensitive data
  - Linking sensitive data

ODISSEI Data Facility Project

Aim

- Provide a secure environment to analyse and process sensitive data on HPC, complying with the legal frameworks such as the CBS law, WGBO and GDPR (AVG).

Scope

- Show that the secure platform can be used to serve various research domains with different requirements such as applications they run and data types being processed. We also show that the environment is scaleable for multiple use cases.

Project Partners

- ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations)
  - Infrastructure for social sciences
  - 32 member instituties
- SURFsara is the Dutch national HPC center and is part of SURF which is a collaborative organization for ICT in Dutch education and research
- Statistics Netherlands (CBS)
  - Social Statistical Datasets (SSD)
  - Registered and administrative data for scientific and social driven research

Solution – A novel approach

- We propose a novel method to create secure computing environment on traditional multi-tenant high-performance computing clusters
- The platform as a service provides a customizable virtualized solution to address the sensitive data challenges without modifying existing HPC infrastructures.
- Using PCOCC and SLURM this platform can be used for processing sensitive data within a shared HPC environment and address both strict and flexible data security requirements.

National Supercomputer Cartesius

- ~ 2000 nodes
- Batch scheduler (SLURM)
- Nodes are connected by a fast low latency interconnect. Called InfiniBand
- Public machine, with many of users on it.
- Not designed to keep data inside the system

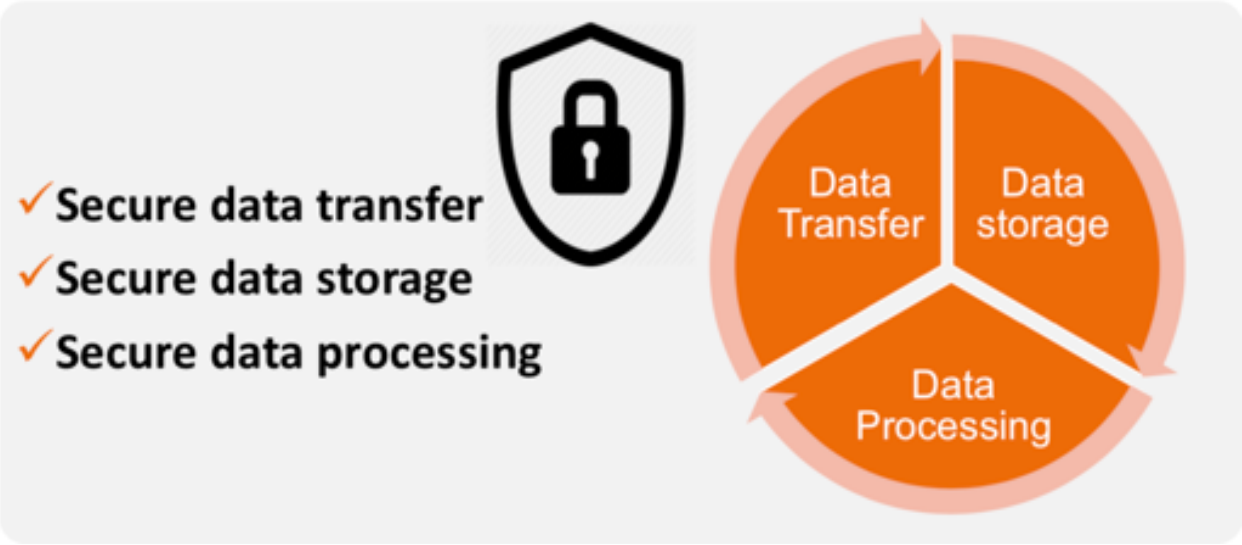
Main technology used in OSSC (ODISSEI Secure Supercomputer)

energie atomique • energies alternatives

- Private Cloud On a Compute Cluster (PCOCC)
- Developed by CEA, French HPC site
- Python
- Opensource software glued together
  - SLURM spank
  - Kvm/qemu
  - Openvswitch
  - Etcd

Secure access and data transfer

- Users can access and transfer data to their private cluster by means of a VPN, seamlessly and securely integrating the cluster into their own private network.
- Stringent automated security controls make sure this VPN is the only path for sensitive data to leave the cluster.



OSSC Pilot showcase

ODISSEI SECURE SUPERCOMPUTER (OSSC)

In the OSSC researchers can analyse their data linked to CBS Microdata in SURFsara's high-performance computing environment. These are three recent projects:

A genome-wide association study of health care costs

Analysing social network of the Netherlands

Effects of spatial contextual characteristics on personal income

Thank you!

Thanks to:

ODISSEI

Open Data Infrastructure for
Social Science and Economic Innovations