EUNIS 2019: Services for Sensitive Data

Gard Thomassen[1]

[1]University of Oslo, University Centre for Information Technology, gardot@usit.uio.no

Keywords
Research services, GDPR, health data, IoT, data collection, data analysis, storage, hpc, IT-security.

1. SUMMARY

Through the Services for Sensitive Data ecosystem (TSD) the University of Oslo delivers a suitable and secure eInfrastructure for nation-wide usage. A wide range of research projects are handling data that according to GDPR are special categories of personal data, and therefore should be handled with special care. The private cloud Platform as a Service today hosts more than 3000 researchers from more than 560 different research projects storing more than 2,5 petabytes of data. TSD facilitates data collection services (web-based, smarthphone/tablet apps and IoT), cross-border collaborations, supercomputing, massive storage and other specialized services. Lately TSD has focused on developing easy-to-use end-user services including easy data import and export, a self-service portal and a digital dynamic consent portal.

2. EXTENDED ABSTRACT

Along with the heavy uptake of research usage of DNA-sequencing machines, video-footage and fMRI imaging within different research groups at the University of Oslo around 2010 the University Centre for IT (USIT) recognized an unmet need of a secure eInfrastructure for handling data containing special categories of personal data, formerly known as sensitive personal data. Additionally, data with high confidentiality requirements in Academic-Corporate collaborations needed such a safe haven.

By bootstrapping resources from ongoing research projects and national and local eInfrastructure providers USIT managed to come up with a secure, multi-tenant Platform as a Service (PaaS) to meet this demand. Initially USIT considered the potential user-group as small, but soon realized that unleashing parts of the potential of such a secure eInfrastructure revealed a large amount of users, and unmet needs. The PaaS system was soon named Services for Sensitive Data (with TSD as the Norwegian acronym). The absolute basic requirements of TSD has been based on the CIA principles; Confidentiality, Integrity and Accessibility). Absolutely all end-user functionality have been based on real end-user demands. Sometimes the end-user does not really know what they want technically, but they do know where they want to go with their research, and what features they need to optimize and realize their work. The list of user-requirements was quite comprehensive at the start, and the list continues to grow:

- Multi-tenant
- Easy collaboration across borders and outside of the home institution
- Unlimited storage with backup
- Supercomputing
- Regular windows and linux computers with normal and specialized research software
- Data collection from web questionnaires
- Data collection from apps and Internet of Things
- BYOD (bring your own device) on the end-user side
- Access from anywhere
- Support for multiple in-sync video-stream playback
- Easy data import and export
- Conditional export privileges
- Database-support

- Low cost
- High availability
- Self service

And the internal list of system operations and architecture requirements was no less of a challenge:

- High level of security (Confidentiality)
- Research project separation (Integrity)
- High availability (A)
- 2-factor login
- Minimum usage of non-personal sys-admin-users
- Granular access even for sys-admin
- As little deviation from regular USIT operations and technology as possible
- Cheap and easy operations
- Self service

During the first two years of operations the major child-diseases have been removed and the system now is stable with a guaranteed availability of 95% per year, with a measured availability of approximately 98% in 2018. The TSD system has seen an immense growth of research projects and totals 560 projects as of 1st of March 2019, see figure 1. The TSD system today has a 2000+ AMD 7551 core high performance computer with 8 GiBs of RAM pr core and additionally three nodes for Artificial Intelligence and Machine Learning algorithms boasting a total of 6 V100 nVidia GPUs. Regarding storage TSD is in the process of storage system renewal and the procurement is ongoing. TSD is aiming at a approximately 4 PiBs of storage during this procurement, and TSD are also renewing the VMWare based virtualization infrastructure.

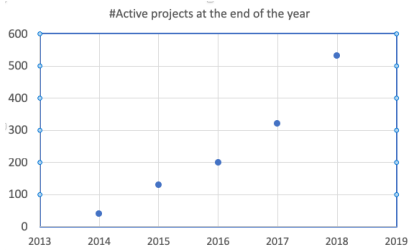

Figure 1, TSD growth pr year

Since 2016 the main focus has been on availability and continued security work, but not at least on automation, self-service and end-user services. The data collection capacity scales freely, and today TSD receives 100.000-150.000 end-user survey replies per week, and the system has been found capable of 100.000 replies per hour without stressing the servers. The data collection services have been an immense success for research usage, and for special cases TSD has programmed smartphone and tablet applications for dedicated usage. Further, TSD has made a secure smartphone Dictaphone and will probably release a similar smartphone video-recorder during 2019. In all such use-cases there is an immense focus on built-in privacy, and in 2018 TSD in combination with the onlkine survey system called “Nettskjema” made second place in the “National championship of innovation, IT and built-in privacy”. All replies are delivered as PGP encrypted files, where each research project has their own key, and the TSD system automatically decrypts and organizes the incoming data for each project each night.

The major development of 2018 was a digital dynamic consent system that enables online digital level 4 (highest public digital security level in Norway) signatures on consent forms, a website where respondents may check, track and revoke their consents. And a portal and API-service for researchers where they have full insight and audit trail of all consents belonging to their project. The system of course opens for manual insertions of paper-consents in the cases where digital signatures is impossible. This is a major step towards build-in privacy and the right to control your own data as stated by the GDPR.

Today TSD is a national eInfrastructure in Norway provided through Uninett Sigma2 AS (the national provider of storage and high performance computing for Norwegian Research use), and operated and developed by the University of Oslo. All major Universities and University Hospitals of Norway have one or more research projects hosted by TSD, and all Norwegian researchers may apply for free or heavily subsidized compute and storage resources through Uninett Sigma2 AS. TSD is even available to EEC researchers as a service on the EOSC-HUB.

3. AUTHORS' BIOGRAPHY

Author Gard Thomassen has an Maser of Science in Computer Science and a PhD in Bioinformatics. Thomassen worked at the Oslo University Hospital as a PostDoc and was part of the team doing the first DNA exome and RNA transcriptome sequencing of tumor and normal samples in Norway om 2010. In 2012 Thomassen started at the University Centre for Information Technology (USIT) at the University of Oslo as project leader for building a system for storage, analysis and collection of personal sensitive data (TSD). These services now represent the national solution for large scale research on sensitive data and hosts more than 2,5 PiB of data, 3000 users in more than 560 research projects, more than 1000VMs and a 2000+ core High Performance Computing cluster. TSD has also delivered the IT infrastructure for clinical deep sequencing at the Oslo University Hospital since 2015. Today, Thomassen is Division Head for Research Computing Services, and Assistant Director at USIT, and he is responsible for supporting a wide range of subjects including Sensitive Data, Digital Humaniora, Statistics, Digital Collections, Mass Storage, Museum IT, AI and High Performance Computing.