1.1 Background

The Development programme for data management and computing infrastructure 2017–2021 (DL2021 programme) was carried on by the Ministry of Education, science and culture together with research and innovation actors. The programme responded to the Research and Innovation Council's vision and roadmap for 2030, especially the goals related to ensuring sufficient competence base, having data and artificial intelligence as drivers of platform economy and increasing the attractiveness of Finland. In the programme, the data management and computing infrastructure of CSC – IT Center for Science Ltd was updated to a level that ensures international collaboration. However, the programme was not only about updating the infrastructure of CSC but also about developing the competences both CSC's and customers' for effective use of the infrastructure and the services. In addition to research, the programme was to support data management and computing infrastructure for education and innovations as well. The national development programme is strongly linked to European research infrastructure policies. CSC was an excellent actor to implement the data management and computing infrastructure due to its competences and international networks but also the location of the data center. This report presents in the light of statistics, user cases and and the actions CSC has accomplished, how CSC has contributed in achieving the goals set for the programme.

Goals for the programme

The programme had goals related both to the infrastructure and services and for competence development.

Goals related to infrastructure and services

- The infrastructure supports research and education.
- With the development programme the infrastucture of CSC – IT Center for Science Ltd. will be developed to the level that ensures the international collaboration.
- The infrastructure and related services will be offered for wider public research community. For governmental research institutes the services are offered under the same conditions than for higher education institutes as of 2018.

Goals related to competence development

- Respond to the new needs for services from the research community and new research fields.
- Enable the management of data and computing so that research and innovation can respond also to new challenges.
- Strengthen the national and facilitate local expert support so that they are able to scale with increasing number of users and entry of research institutes.

Use cases

There were six main use cases and research needs identified as the drivers for the programme. The programme, both by upgrading the infrastructure and developing competences were to response to the needs.

Large-scale simulations

- Traditional high-performance computing, that benefits especially physical sciences
- Research topics include climate change, space weather, fusion reaction, phenomena of astronomy and particle physics

Medium-scale simulation

- Material sciences research, energy-technology challenges, research questions in chemistry and other natural sciences (e.g. biophysical simulation of cell behaviour)
- Utilisation of geospatial data

Data-intensive computing

- Bioinformatics research
- Language research and other digital humanities
- Use of data-analytics in analysing business data and economics

Data-intensive computing on sensitive data

- Medical research based on patient records, e.g. cancer and epileptics research and analysis of genome and imaging data
- Interview data of social sciences and humanities and voice, image and video recordings
- Registry data and other identifying personal data
- Requires among other things higher security level environment and tools to authorise the data use

Artificial intelligence

- Learning algorithms and utilisation of diverse data sources fo challenges in science and business, e.g. artificial vision and intelligent traffic
- Usage is expanding to other areas, as bio sciences and humanities, e.g. natural language research

Internet of Things (IoT) and data streams

- Utilisation of continuous data streams, as measuring satellites, weather radars, sensor networks, stock prices, social media message streams
- Utilisation of sources of data stream connected to the internet (Internet of things), e.g. in robotics and applications in industry

---

1.2 Execution of the programme

Before DL2021, the previous computing infrastructure of CSC was from 2011. Digitalisation and growth of data intensive research brings new requirements for computing infrastructure. Handling large amount of data requires more memory and storage capacity than traditional high-performance computing. The upgrade for the data management and computing infrastructure was designed based on the needs, also brought up by the users of CSC's services. Safe and easy-to-use systems enabling transferring and storing large amount of data, data analytics and data-intensive computing was emphasised. To enable smooth transition from the old system to the new one, and avoiding long periods of the services being unavailable to the users, the installation was conducted during several years. The installation and getting the services into production was delayed from the original plans for reasons not dependent on CSC. The delays did not cause any major consequences for the customers' work. The largest cut on the service was when Sisu was decommissioned and Puhti was not in production yet. However, temporary computing capacity was offered by a collaborator organisation from Edinburgh, Scotland.

Investment

The programme was funded during 2017–2021 with the following investments:

- Competence development and new services and support for new user groups and new use cases (2 M€)
- Development of DL2021 data management and computing environment (33 M€)
- Additional funding for increased AI capacity (4 M€)

Main components of the infrastructure are briefly presented below with links to further information in CSC's webpages.

- Puhti allows customers to run serial and small to medium sized parallel jobs through a batch queueing system. Puhti includes CPU based nodes with a range of memory, and nodes with Nvidia Volta GPUs (Puhti-AI) for HPC and AI workloads.
- Mahti is CSC's flagship supercomputer with a peak performance of 9,5 Petaflops. Mahti has 1404 CPU nodes and is meant for larger jobs (minimum 128 CPU- cores). Mahti-AI includes 24 GPU nodes based on Nvidia Ampere A100 GPUs.
- Allas is a general-purpose data storage service for storing and sharing data, based on CEPH object storage technology. The stored data can be made accessible over https via a public URL. Data processing can be done using standard APIs from anywhere. Allas – in production September 2019

The DL2021 environment was further diversified in 2020 when Kvasi, a quantum computation simulator, was introduced. Kvasi enables the simulation of algorithms on quantum computers of up to 30 qubits.

The services for sensitive data management, that utilises Allas and ePouta, have been developed partly with the DL2021 funding. Two of the services, SD Connect and SD Desktop were opened for public beta use in the first half of 2021.

Webpage link: 1.1 CSC’s datacenter environment, DL2021 data management and computing infrastructure.

Competence building

In order to ensure effective use of the infrastructure, competences for the usage were enhanced through different means. Different training events and a roadshow were organised and in addition to regular expert and customer service, an on-site support was offered for the governmental research institutes. T hese are presented in more detail in section 4 Support and training.

As the programme also opened the use of the infrastructure and services for governmental research institutes it was important to ensure the from the start they are in effective use, both by individual organisations and also in collaborations. This was ensured by pilot projects with several governmental research institutes.

Automatisation of large data streams

- Aim of the project was to analyse thee requirements to automatise the data management of devices that produces large amount of data
- Use case was the data management requirements of genome sequencer by Institute for Molecular Medicine Finland and Finnish Food Agency but the results of the project are applicable to other cases
- Project focused on safe data transfer from the customer organisation to DL2021, processing the data according to the customers' requirements and distributing the products for authorised users.
- Project influenced development of instructions for storing sensitive data in Allas.
- Technical solutions are part of Allas's data management layer.

Secure remote desktop for sensitive data management

- Aim of the project was to produce a secure solution for customers who don't want to use CSC's ePouta cloud service in their organisations' intranet
- Identified user gets access to data in the cloud service to which the owner of the data has granted permission.
- Related also to the cooperation with Findata
- CSC's new sensitive data services, Sensitive Data Desktop and Sensitive Data Connect, were opened for public beta use in 2021

Datasets as a Service concept

- Higher level concept considering several amendments in CSC's environment
- Aim of the project is to enable offering data from CSC's environment for broader use, including a solution for data staging, being able to use Allas data in Puhti.
- In addition to research data, one major use case being explored was offering companies' datasets for research use

Automatised IoT datastreams

- Relevant for example for managing and processing data from wearables or from weather stations
- Solution based on Apache Kafka and Apache Spark technologies
- Ready-made solution to Rahti environment

Next chapter: 2. Infrastructure benefitting the whole research and education community