Data Science Storage

Welcome to the LRZ Data Science Storage documentation.

Overview

LRZ's Data Science Storage (DSS) is a novel approach at LRZ to solve the demands and requirements of data intensive science. Therefore, DSS implements a data centric management approach, which gives you the ability to:

- Store vast amounts of data for as long as the data is important to you or the science community
- Access this data from the whole LRZ computing ecosystem (SuperMUC, LinuxCluster, Compute Cloud, VMWare, Remote Visualization, Housed Customer Compute Systems)
- Share this data between arbitrary users of the LRZ computing ecosystem
- Access/Transfer this data world wide via a high performance, WAN optimized transfer protocol, using a simple Graphical User Interface in the Web
- Share this data with arbitrary users around the globe, like you are already used to from services like LRZ Sync+Share, Dropbox or Google Drive

Who can use it

Currently DSS is only available as a so called joint project offer from LRZ. This means that we will analyze your requirements together and then provide you with an offer for purchasing, implementing and housing a dedicated DSS building block for 5 years that exactly fits your needs. The offer consists of a one time investment part, for which you usually can apply for funding by DFG or other funding agencies, and a yearly service fee, which covers the costs for managing and running the storage system at LRZ. Economically attractive configurations for this model start at around 0.5 PBs.

In addition to that, we are currently in the process of applying for funding of our own DSS building block, for which we intent to provide a certain amount of storage as a base service for our HPC user base, for as long as we get central funding for it. However, the terms and conditions that will apply to this service have not yet been worked out completely.

How does it work

Management

At the very top level of the DSS data management approach are so called data projects. These data projects are special LRZ projects which are used as an organizational envelope of the data, associated with this project. Each data project is assigned grants from one or more data pools.

Additionally, each data project contains a set of users, called data curators. The data curators manage, own and take responsibility for the data, stored within the data project. Typically the principal investigator (PI) of a research project and one or more proxies own the role of the data curator.

In the context of a data project and within the limits of its grants, the data curators create one or more so called containers. Within the containers, a user can store data. In addition to that, containers have certain properties like:

- who is allowed to access the container
- how these access rights are enforced within the container
- how much data can be stored in the container
- how much data a specific user can store in the container
- to which Cloud or VMWare machines the container is exported
- if data in the container should be protected by regular tape backups or even archives

These properties are managed by the data curators using the DSSWeb Web Portal.

Access to data

A DSS container is basically a special directory on a shared file system (the data pool). Associated with this directory are one or more user groups, that define the users which can access the container and an enforcement policy. The enforcement policy defines how the access rights, defined by the data curators on the container level, are enforced on the data placed within the container.

On SuperMUC and the LinuxCluster, all DSS file systems are currently available on the Login Nodes via the path /dss/. For virtual machines in the Compute Cloud or the LRZ VMWare infrastructure, data curators can export their containers via NFS to these machines.

Outside of the LRZ compute ecosystem, DSS containers can be accessed and shared via the GridFTP protocol. The most convenient way to do this, is to use the Globus Data Management Web App.
**Infrastructure**

On the infrastructure level, DSS consists of multiple storage building blocks, that are integrated to a single virtual storage system, using the IBM Spectrum Scale software defined storage stack. This software allows us to scale DSS to hundreds of PBs, if required. The storage building blocks are connected to a dedicated high performance interconnect, which provides high speed connections to all of LRZ's compute systems.

**Where to go from here**

- If you are interested in our DSS project offering, please contact us via the LRZ Servicedesk (Datenhaltung Data Science Storage).
- If you are a principal investigator, looking for information how to start a data project, continue reading here.
- If you are a data curator of a data project, looking for information how to manage containers and data within a data project, continue reading here.
- If you are a user, invited to access a container, looking for information how to use DSS, continue reading here.

**Where to get further help**

If you have a questions regarding DSS that has not been answered by our documentation or encountered a problem with DSS, please contact us via the LRZ Servicedesk and open a ticket for the Service Datenhaltung Data Science Storage.