
Data management

Overview

When realising complex scientific applications or industrial workflows, data management is one of the crucial ingredients. Consequently, UNICORE integrates powerful data and metadata management features.



Storage backends

The UNICORE SMS (storage management service) provides an abstract interface to a variety of backends. Currently, the following backends are supported:

- File systems can be accessed via the UNICORE Target System Interface (TSI). This is the default storage backend.
- S3: cloud storage accessible via the S3 interface (e.g. Amazon, OpenStack, ...). UNICORE can flexibly manage the required access keys.
- Apache Hadoop HDFS: the popular HDFS file system can be accessed via UNICORE as well, making it possible to combine Hadoop and HPC into application workflows.

Other backends can be added, as the implementations are pluggable.

Data transfer

The ability to move data efficiently from client to server, or between servers, is realised using file transfer services.

- Multiple protocols: depending on the storage back end, different file transfer protocols are available, and can be used transparently.
- Server-server transfers: when copying data from one server to another, the start time of the transfer can be controlled. Server-server transfers are managed by the server, so the client can go offline.
- High performance data transfer: the UNICORE file transfer protocol (UFTP) combines high performance with firewall-friendly configuration. Data connections are dynamically allocated and managed. Additional features include multiple TCP streams per transfer, optional data encryption and data compression, allowing to fine-tune UFTP to your requirements.

Metadata management

UNICORE has a built-in metadata system with a number of distinct features.

- Metadata is stored in hidden files right next to the data, which allows to keep data and metadata consistent and making backup and archiving easy.
- Automated extraction: UNICORE can automatically retrieve and index metadata from your files using the Apache Tika extraction framework.
- Search engine: indexing and querying the metadata using Apache Lucene, a powerful open-source Java search engine.
- Search only one given storage, or search multiple storages across the UNICORE federation.

Data-driven processing

UNICORE can be configured to automatically process files according to a set of user-defined rules. It's possible to run short scripts, batch jobs or to extract metadata.

Further Reading

www.unicore.eu/documentation/architecture