RDA Data Security/Resilience Overview

Data Preservation

- The Computational and Information Systems Lab's (CISL) maintains strategies to ensure access to and availability of data holdings for three- to five-years as well as long-term. The strategies include activities shown below.
    - 24x7 monitoring of hardware and software for warnings and errors.
    - Support to create and maintain disaster recovery data copies as described here.
    - Proactive replacement of marginal tapes.
    - Data migrations from older media to new media.
    - Periodic replacement of older hardware.
- A unique copy of each archived data file is maintained separately on disk and tape media storage. A third copy is maintained in the disaster recovery tape pool for datasets that have no copy at a trusted alternative repository, such as the Copernicus Climate Data Store.
    - The primary copy is hosted on the glade file system.
    - The backup copy is hosted on the Quasar tape system.
    - Disaster recovery copies are written to Quasar and the tapes are removed and stored in a fireproof safe at the NCAR-Wyoming Supercomputing Center (NWSC) facility located in Cheyenne, WY.
- RDA Web Server Disk directories that host datasets' metadata are backed using a Veritas Backup Exec system on a daily basis to storage servers at both the NWSC and NCAR Mesa Lab (NCAR-ML) facility in Boulder, CO. This creates a geographically separated backup copy from the servers that are hosted at the NWSC.
- RDA Metadata Databases, which include repository inventory information and file level metadata, are replicated on MySQL database servers running on separate physical hosts at the NWSC. Additionally, the full databases are backed up from the NWSC servers to NWSC and NCAR-ML hosted storage servers on a daily basis using the Veritas Backup Exec system described above.

Data Security

- Checksums are computed on all archived data files; i.e. the files that are stored on the RDA Dataset Collection Disk, and Quasar Tape backup systems. The checksum information is recorded in the RDA Metadata Database and is used to verify file integrity. Historical checksums are verified during the following operations:
    - When files are restored from tape to disk for any operation.
    - Random checks of a sample disk and tape files on a weekly basis.

  If a mismatch in checksum is discovered on any of the stored files, a report is created and a staff member manually checks to verify that a second copy of the file on tape or disk matches the recorded checksum, and subsequently can be used to replace the

corrupted copy. If needed, data may also be downloaded from an alternate trusted repository to replace a corrupted data file.

- Data are protected from being overwritten by unauthorized parties through the use of Portable Operating System Interface (POSIX) file user and group settings on RDA Dataset Collection Disk and Globus user permission settings on Quasar.
- The RDA requires its users to register in order to access RDA datasets. Users are required to authenticate on CISL High-Performance Computing (HPC) systems or through the RDA web server to get read access permission to all archived data files stored on RDA Dataset Collection Disk. Non-RDA staff users are only allowed read access permission in all cases. Authentication logs are monitored to detect repeated attempts to unsuccessfully authenticate.
- Operating systems and related software are updated to the most recent versions on a monthly basis, or as needed more frequently to ensure that security patches are current.

Risk Management and Resiliency

- CISL staff, including those that manage the RDA, participate in institutionally led Enterprise Risk Management (ERM) and Risk Assessment exercises to identify and manage risks on a regular basis. Additional information on UCAR's ERM office is provided at https://rda.ucar.edu/rdadocs/UCAR-Enterprise-Risk-Management.pdf. A description of UCAR's ERM framework is provided at https://rda.ucar.edu/rdadocs/ucar_erm_framework_v1.0_june_2019.pdf.
- Dataset metadata, documents and software are backed up using a Veritas Backup Exec system on a daily basis to storage servers at the both the NWSC and NCAR-ML facilities. This creates a geographically separated backup copy from the servers that are hosted at the NWSC. Using this strategy, three independent copies of this information always exist.
    - One copy on NWSC hosted RDA Web Server Disk
    - One copy on NWSC hosted backup storage server
    - One copy on NCAR-ML hosted backup storage server
- Database replication across physical hosts ensures two physically separated nearly real-time copies of RDA Metadata Databases exist at all times. All RDA Metadata Databases are systematically backed up on a daily basis using the Veritas Backup Exec system described above.
    - One copy on production NWSC hosted RDA Metadata Database
    - One copy on replication NWSC hosted RDA Metadata Database
    - One copy on NWSC hosted backup storage server
    - One copy on NCAR-ML hosted backup storage server
- All bin/log database logs are backed up once per week to Quasar tape storage.
- An in-house developed machine learning tool is in place to detect and alert about anomalous file remove patterns.
- Proactive relationships with colleagues, vendors, and industry are maintained to track technology trends.

Maintaining Storage Media

- CISL monitors tapes for warning and error messages, and if a tape is detected to be marginal, data are copied from the tape.
- Quasar tape media verification is enabled and monitored by CISL.
- CISL upgrades to the latest tape technologies on a 3-5 year basis, and migrates data from older tapes to the new tapes.
    - Furthermore, the value of the RDA is recognized, so that as a priority for NCAR, the RDA datasets are stored on a specific set of tapes in order to make it easy to systematically migrate the tapes to new technology.

Disaster Recovery

- The reference copy for all RDA's archived data files is stored on disk and tape media storage at NWSC. A third copy is maintained in the disaster recovery tape pool for datasets that have no copy at a trusted alternative repository and stored in a fireproof safe at NWSC.
    - If either the disk or tape media storage experiences a major failure, the copy hosted at the on the unimpacted storage system can be used to repopulate the other system once it is restored. A detailed plan specifies the process to recover data lost on disk systems from the copies hosted on the Quasar tape system.
        - https://rda.ucar.edu/rdadocs/RDA_Strategy_to_recover_from_Major_disk_failure.pdf
    - If there is a major disaster at the NWSC facility that disables both the disk or tape media data copies, irreplaceable data will be recovered from disaster recovery tape media hosted in the fireproof safe at NWSC, and from trusted alternate repositories, such as the Copernicus Climate Data Store, which are tracked in RDA dataset metadata. RDA dataset metadata would be restored from the backup copies hosted on NCAR-ML storage by CISL staff
- All metadata, databases, documentation, and software associated with the RDA's datasets can be restored from the Veritas Backup Exec system by CISL staff when needed.
- Support contracts are maintained with IBM that enables CISL HPCD to send tapes with unreadable data blocks to the vendor for recovery.
- Support contracts are maintained with DDN that provide a high level of support for RDA Data Collection Disk. The vendor can be called upon to repair disk related hardware failures and file system contents when needed.