10.11 DATA ACCESS AND STORAGE IN THE LEAD CYBERINFRASTRUCTURE

Anne Wilson*, Doug Lindholm and Tom Baltzer
Unidata Program Center, UCAR

1. Introduction

The Linked Environments for Atmospheric Discovery (LEAD) project, funded by the National Science Foundation, is building a cyberinfrastructure for mesoscale meteorology research and education. In LEAD users can build and execute orchestrations involving multiple data sources and the web service-based tools provided by LEAD. LEAD tools and applications include data mining algorithms such as Algorithm Development and Mining System (ADaM), assimilation tools such as ARPS Data Assimilation System (ADAS), and forecast models such as the Weather Research and Forecasting (WRF) model.

These orchestrations generally require input data and generate both intermediate and final data products. The orchestration must access and likely move the input data and must have storage space to store resulting data.

A major goal of LEAD is to allow users to query, import, and manage data in the LEAD domain for purposes such as visualization, storage, and as input to LEAD applications. Data used in LEAD may be stored in the LEAD Data Repository, or may be data provided by cooperating institutions, or may be other outside data known to the user. LEAD data can be individual files, collections of files, or streams. LEAD data sets may be very large, necessitating storage in a distributed manner or on a mass storage system. These large sizes also require support for subsetting, subsampling, and possibly aggregation of data.

This paper discusses data access and storage by both users and orchestrations in the LEAD cyberinfrastructure. We describe the integral role of metadata in data access and characterize metadata generation. Finally, we provide a canonical use case to demonstrate these concepts.

Corresponding author address: Anne Wilson, UCAR/Unidata, P.O. Box 3000, Boulder, CO 80307; email: anne@unidata.ucar.edu

2. The Role of Metadata

Metadata plays a critical role in LEAD data access. By providing a handle to the data, metadata provides location independence of the data. Additionally, metadata is necessary to support queries such as, "I want the most recent day's worth of radar data over Oklahoma." Metadata can also support data browsing, a helpful activity for student learning through data exploration.

Thus, metadata must be generated for any new data products brought into the LEAD sandbox. The LEAD metadata schema is based on the FGDC metadata schema. The minimum metadata requirements for LEAD include platform, a temporal range, a spatial range, and identification of the variables where appropriate.

Under most conditions metadata can be harvested automatically. Public LEAD data (described below) has THREDDS or some other compatible metadata available, which is transcribed into the LEAD metadata schema via a crosswalk process. For data that has no metadata, sometimes that metadata can be generated by examining the file name and/or its contents. For example, netCDF files can contain name and attribute information for variables as well as other information such as source and creation date. Also, orchestrations in LEAD, the programs that coordinate data staging and application invocation, have information about new data products being generated that can be used in metadata generation. However, in the case where the minimum metadata can not be generated automatically, user interaction will be required.

Metadata generation will be facilitated by use of the Unidata Common Data Model (CDM). The CDM provides a unified interface to a variety of data formats by virtue of an abstract data model, which allows uniform access through a set of standard APIs.

Real time streaming data requires appropriately constructed metadata to accommodate the dynamic nature of the data. It would be overly cumbersome to create and

manage metadata for every real time data product originating from a stream. Instead, streaming data will be cataloged as a collection and thus the metadata description will remain static over the lifetime of the stream. The spatial range described in the metadata will be broad enough to cover the spatial range of the entire stream. The temporal range will reflect a time period from "now" to some time in the past that reflects the configuration of the local space management policy.

3. Storage Classes in LEAD

There are three classes of data that the LEAD data storage subsystem must handle. Personal data is data that a user has brought into their space within the LEAD Data Repository. Users may store data and other resources here. Also the orchestration will store intermediate results in a user's personal space at their request. Public LEAD data is data that is made available to the LEAD community by cooperating data providers, such as universities or other institutions interested in sharing the data that they receive or generate. Users can discover public LEAD data via querying the LEAD Resource Catalog. External data is any other data that a LEAD user knows about and would like to access, such as NCDC archival data.

LEAD plans to support data access via OPeNDAP, gridftp, ftp, http protocols and possibly others. OPeNDAP in particular supports subsetting and subsampling, valuable features in transmission and management of large data sets.

4. Personal Data

Storage space within LEAD is called the LEAD Data Repository (LDR). Individual users may use this space for short term storage, such as the staging of input data, or for longer term storage. Additionally, an orchestration can be configured to save all input, intermediate, and output products to the user's personal.

The access, privacy and security guarantees of this space are comparable to that of a user's local file system. Similarly, the user is responsible for ensuring that she stays within her space quota in the Repository. Supporting tools can help with space management, such as a scouring utility to aid in management of streaming data. Presumably the user will save especially interesting data in other locations that are not scoured.

The LDR is catalogued via the myLEAD cataloging system, so all metadata-based functionality such as query and ontology support is available.

Importing data into the Repository may invoke a metadata generation routine if the minimum LEAD metadata does not yet exist, which may require user input in order the meet the LEAD minimum metadata requirements. In particular, setting up the importation of streaming data will require an interactive session to define metadata fields. Removing the stream from the repository will similarly require that myLEAD be appropriately updated.

The LDR is currently being implemented using the Unidata THREDDS Data Repository (TDR). The initial implementation is distributed across a variety of LEAD test beds that reside at several LEAD institutions, such as Unidata, University of Alabama, Huntsville, and the University of Indiana. Later implementations are expected to include mass storage devices.

To ease the handling of large data sets, the TDR will ultimately support subsetting and subsampling, again through the Unidata Common Data Model (CDM).

Ultimately, robustness and reliability in the LDR will be achieved via replication.

5. Public LEAD Data

Public LEAD data consists of data acquired and managed by cooperating institutions outside of LEAD that provide support for access to that data. Currently the majority of public LEAD data is streaming data acquired by sites via the Unidata Internet Data Distribution (IDD) network, but some assimilation data generated by the Center for Analysis and Prediction of Storms (CAPS) is also available. Soon some NCDC model data will also be available as public LEAD data.

For LEAD users and orchestrations to access this data, the data provider must provide metadata that is compatible with the LEAD the metadata schema, that is, a crosswalk must exist between the two.

Currently, all public LEAD data is catalogued via the Unidata THREDDS metadata schema, and in some cases via a Unidata THREDDS Data Server (TDS). Data providers must also provide a URL and supporting functionality to access the data such as running an OPeNDAP server or making it available via gridftp.

LEAD maintains a global catalog of metadata for data and other LEAD resources, called the Resource Catalog. Public LEAD data providers can request that the Resource Catalog be configured to catalog their data, which allows

LEAD users to discover their data via the query interface.

LEAD can control access to the metadata for public LEAD data, but not the data itself. Data providers will be able to specify an access policy to the metadata that is enforced by the LEAD Resource Catalog to ensure that only permitted groups can access or query over the metadata. This is independent of the access policy of the data itself, which is determined by the data provider, although both these policies should be consistent with each other.

It must be understood by LEAD users that these data may be unavailable at times due to site policies and/or technical difficulties. However, it is common for multiple sites to receive the same data sets, providing an informal form of reliability and robustness through duplication. These data providers also determine their own space management policy. It is expected that these data will have a limited lifetime, so this is considered a short to medium term storage.

Most public LEAD data will be grouped and cataloged as collections, to mitigate the expense of cataloging every individual data product.

6. External Data

LEAD plans to support access to other data that is not being cataloged via the Resource Catalog. Thus, this is data that a user would need to know about a priori. While some of this data may be available quickly and programmatically, it may also be the case that staging this data for importation may take hours, days, or even longer depending on the space requirements of the data set and whether human intervention is required to stage the data.

Much public access to atmospheric data is for interactive purposes such as visualization and thus may be in formats that are unsuitable for some LEAD applications. But, as recognition of the utility of raw data grows, efforts are similarly growing to provide unformatted access to raw binary data.

Data access provided to the public is often provided via interactive web pages. When staging of the data is required, availability notifications may be sent via email, requiring the user to provide and monitor a valid email address. Or, a LEAD user may wish to import data provided by a colleague via ftp or gridftp. Although these interactive characteristics can be hurdles to a building a totally automated system, the LEAD portal could provide support for LEAD users to select, request, retrieve, and store this data to a location accessible to LEAD applications. This importation process may invoke an interactive metadata generation process to generate metadata fields that can't be harvested automatically.

As an example of external data, NCDC and collaborators provide access to model data via NOMADS, NOAA Operational Model Archive and Distribution System. NOMADS is an interagency agreement to provide data access via OPeNDAP. NOMADS also provides ascii access to data via ftp and http. Some of this data is old on the order of years. NCDC is thus a truly long term archive that can provide reliable data access to LEAD users such that they need not store copies of the data elsewhere.

As an example of the additional vagaries of handling external data, while NOMADS serves data via OPeNDAP, not all the data are immediately available on line. Older data is stored off line and must be staged before it is accessible. Additionally, off line data may not be subsetted because it is stored in compressed format. This limitation may restrict staging of a large request due to space limitations.

7. A Canonical Use Case

To demonstrate the use of the three classes of storage consider the following scenario, illustrated in Figure 1. A user wants to run regularly scheduled, high resolution, steered WRF forecasts over her state. She will use a data mining algorithm to determine the areas of greatest precipitation, which will determine where to focus the model.

She uses the LEAD query interface, which queries over public LEAD data, to discover a source for near real time level II radar data within her state. She uses the orchestration tool to create an ongoing schedule of importing the radar data via LEAD supported data movement technologies (such as OPeNDAP or gridftp) into her personal space to ensure that each iteration of the orchestration will have the most recent radar data available. LEAD metadata for these radar data are generated automatically from the existing THREDDS metadata.

She also wants to mine her local mesonet data, which is not cataloged by the LEAD Resource Catalog, and is thus external data. This data is updated periodically and served via ftp from a remote server. Using the LEAD portal she sets up a regularly scheduled ftp action to bring the mesonet data into her personal space in the LEAD Data Repository. As part of this

importation process, she specifies metadata values that are used in a metadata generation process for each product retrieved. Some of these values are expressions. For example she defines an expression that uses the product file name to generate a time range. She then uses the orchestration tool to also feed this data into a data mining tool built to handle this mesonet data.

The data mining tools each determine a region of interest (ROI). These regions are fed to a region resolving process that determines a final ROI. This resulting region of interest is one of the inputs into the data assimilation tool, along with other observations and model output.

restarting after a failure) she must compose her orchestration using tools that are components of ADAS.

Now that the user has constructed her orchestration, a few details remain before launching it. First she must configure the orchestration to loop indefinitely. The scheduled importation of both the radar and the mesonet data ensure that the most recent input data is available in her space in the Repository. Finally, via the myLEAD cataloging interface she is able to indicate that all data products generated as output from the WRF tool will be designated as sharable to all other LEAD users so that others may view the results by pointing a visualization tool to the resulting data stored in her personal space.

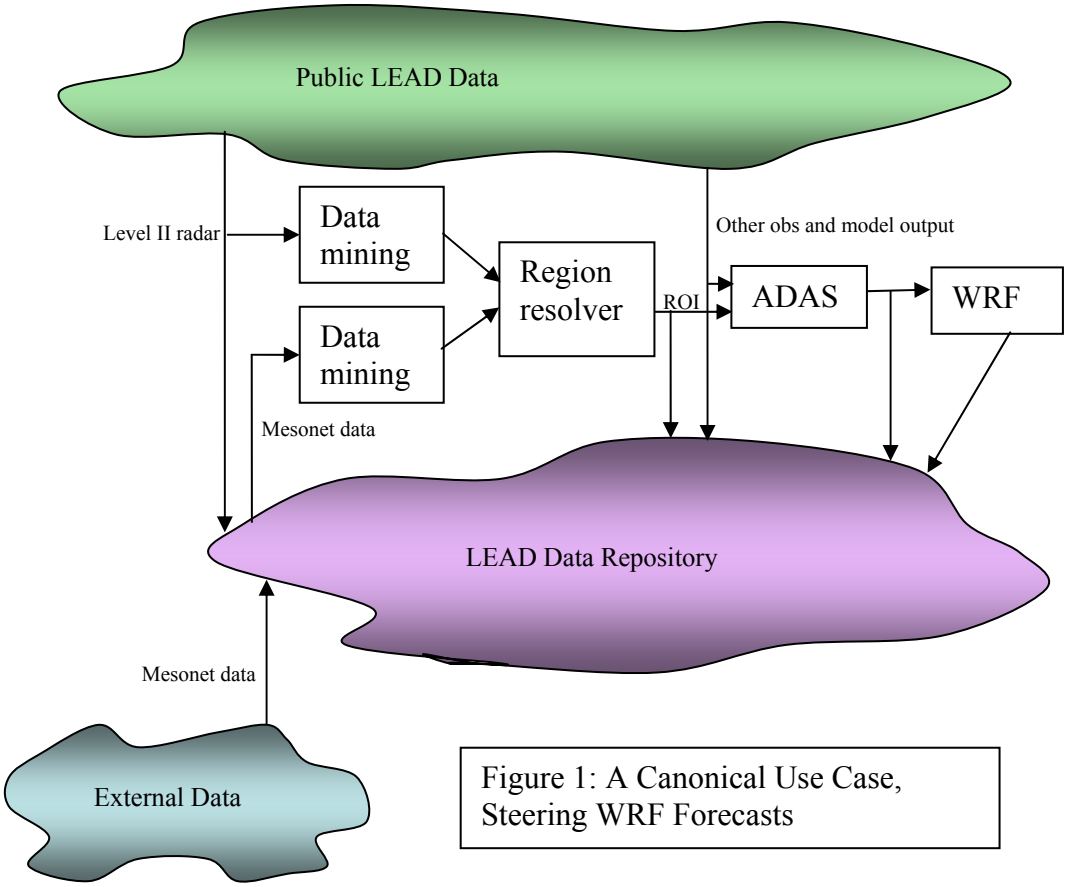

Figure 1: A Canonical Use Case, Steering WRF Forecasts

The output of the assimilation tool is fed into the WRF tool which generates a forecast. Note that by virtue of using the ADAS assimilation tool as a logical unit, intermediate files generated within ADAS are not saved, as the intent is to shield the user from implementation issues such as these. If an experienced user does wish to save such files (for example to ease the burden of

8. Conclusion

Access to and use of data is a pillar in the foundation of research and education in atmospheric sciences. LEAD's ambitious goals are to provide access as wide as possible to available data, include data both within and outside LEAD. Our hope is that via reliable and robust data access and handling, LEAD users will

be able to store, visualize, apply tools, and share data reliably and with ease.

9. Acknowledgements

Many thanks are due to the participants of the LEAD data thrust group for their contributions and support in this effort.

LEAD is funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594, ATM-0331591, ATM-0331574, ATM-0331480, ATM-0331579, ATM03-31586, ATM-0331587, and ATM-0331578.

10. References

Caron, J., Davis, E. R., Ho., Y., and Kambic, R. P., 2006, Unidata's THREDDS data server, *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, Atlanta, Georgia.

Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S., Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL, *Journal of Digital Information*, Volume 2 Issue 4, Article No. 114, 2002-05-29.

Lindholm, D., Wilson, A., and Baltzer, T., 2006, An architecture for the LEAD data repository, *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology,* Atlanta, Georgia.

NOMADS: http://nomads.ncdc.noaa.gov.

OPeNDAP: http://www.opendap.org.

Plale, B., Ramachandran, R., and Tanner, S., 2006, Capabilities of the LEAD Data Subsystem for Enhanced On-Demand Mesoscale Meteorology Research and Education, *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, Atlanta, Georgia.

Unidata Common Data Model: http://www.unidata.ucar.edu/software/netcdf/CDM/index.html.

Unidata Internet Data Distribution System: http://www.unidata.ucar.edu/software/idd.

Unidata Local Data Manager: http://www.unidata.ucar.edu/software/ldm.

Unidata THREDDS Catalog Specification: http://www.unidata.ucar.edu/projects/THREDDS/tech/catalog/Inv/catalogSpec.html.

Unidata THREDDS Data Server: http://www.unidata.ucar.edu/projects/THREDDS/tech/#TDS.