The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
March 4, 2020, 12-2 pm ET
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

Misha Ahmadian (TTU)	David Martin (ANL)
Richard Carlson (DOE/SC	Lavanya Ramakrishnan (LBL)
Vipin Chaudhary (NSF)*	Hakizumwami Runesha (UChicago)
Melissa Cragin (SDSC)	Arjun Shankar (ORLN)
Sharon Broude Geva (UMichigan)	Alan Sill (TTU)
Dan Gunter (LBL)	Suhas Somnath (ORNL)
Zack Ives (UPenn)	Richard Wagner (Globus)
Joyce Lee (NCO)*	Sean Wilkinson (ORNL)
Jeremy Leipzig (Drexel CCI)	

Proceedings
This meeting was chaired by Richard Carlson (DOE/SC) and Vipin Chaudhary (NSF).

Guest Speaker: Data Integrity (Session 2)
Richard Wagner, Professional Services Manager, Globus, *Data Integrity Management with Globus & BDBags*

Topics (Slide 1):
Globus: Reliability: getting same data to desired destinations.
BD Bags: Tracking validity and integrity of data: provides simple functionality that is reusable in different contexts.

Examples: E3SM to CMIP6 Process (Slide 2-3, diagram)
Earth Science Climate Modeling project: Large scale simulations on DOE leadership systems;
Publish data in ESGF (Earth System Grid Federation), distributed globally and made available to researchers. Currently archived at NERSC.

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

- Data publication process leads to eventual aggregation (hundreds of TB on many different systems)
- Need to focus on data integrity. Silent data corruption - File systems, which are built on networks, and networks, eat data. Data breaks down.

Goal: From point of tape archive to publication, ensure valid and correct data is produced. Try to detect corruption as early as possible.

Globus Value Proposition (Slide 4 -6, diagram)

Supports fast and reliable data transfer directly from your storage systems.

1) Copying data between distributed systems across organizational boundaries. Reliably copying files between arbitrary storage systems across arbitrary communication channels to another arbitrary system
    a. Focus on speed and addressing storage, network failures
2) Verify data after copy or transfer/verify file integrity; if not match, need to re-transfer file

Now supports BYO Checksum capability: Can provide external check – not have to rely on existing data sources as point of truth:

- Ensures that no longer have to rely on existing file in file system as point of truth
- Given manifest or archive report of data, now can talk to endpoints in 2 different centers to verify data
- Intended for data repositories: can report Checksums and set specification
- Currently in API and will go to CLI next

Example: E3SM Data Stager (Slide 7-8)

Ensure data gets to a destination with more confidence. Data stager script:

- Builds on zstash archive tool. When storing data in E3SM, generates and stores checksums for files that helps to package data. Checksums can validate data not lost in process of storage and extraction from archive
- Run data stager: zstash helps generate file manifest with Checksum. Stages needed file on NERSC file system. Submits transfer, including manifest as one of files to be moved. Publication process ensures manifest with data. Can check data on systems while processing is ongoing.
- Given need to manage manifest, ensures needed steps are taken consistently

BD Bag (Slide 9)

Manifests:

- Simple profile based on Bagit standard - Library of Congress uses to package or encapsulate data into zip file.

Bagit/ BD Bag:

- Recommended for getting data around; good if have to bundle files and want to ensure integrity

BD Bag profile (Slide 10)

- Test file concept: can tell client tool how to retrieve data and validate data expectations
- Data directory may be empty. BD Bag then becomes a way to have manifest and describe data to encapsulate
- BD CLI – supports various communications, identifies re-direction, etc. (Slide 12)
- Validate & Materialize (Slide 13)
    - Validate: if unroll bag on system, can run validate flag (check it)
    - Materialize. Test file. Data management becomes simplified. Makes accessible.

Globus Use case**:** When user-select creates data set, create BD bag with set of references and meta data used; way of assembling ad hoc data sets

NIH Metadata Ingest (slide 14, diagram)
Receive metadata in BD Bag and upload to repository, which is 1) staging area for ingest load of catalogues on right; 2) since prominent and simple format, can use as redistribution. If BD Bags intended to be for data release collections or versioning, they are in repository. If want to do more than what catalogue does, research can go back to get base data.

BD Bags fully materialized. Becomes unit of encapsulation that can serve various roles.

- For a portal encapsulating a collection for a user so they can download it, or we have services if bag has references to Globus endpoints, can transfer around

Can use for organization – data goes in data director; can come from different sources and be reorganized. Encapsulate logical organization and means to validate data.'

References (Slide 15)

- Globus Transfer Docs (checksum mechanism)
- E3SM Data Stager (Example of script that grabs files, transfer and put in manifest)
- Bagit & BDBag (Tribal spaces – if want client tool to do well with BD Bags, including materialize or validate)

Discussion
BD Bag addresses data loss: that could be accidentally introduced. Also, security issue: protecting data integrity from bad actors.

- Response: In Globus space of data transfer, have validation step and sometimes optional encryption. If have trusted system with set of Checksums, and it submits transfer to a Globus endpoint and mandates encryption and internal Checksum –> basically, conducting doing more risk mitigation by specifying your trusted Checksum from your trusted system to data between systems that may be outside of your control.
  Faith in Globus to enforce your checksum. Thus, adding additional control to mitigate - so first system no longer has to be point of truth for integrity of the file

Provenance Project started with monitoring I/O performance. Now looking at data security in multiprocessing systems. Tracks anything in cluster touching a dataset and creates a non-reputable record of every process and I/O operation that took place. Complementary aspects of data integrity issue from security point of view. Here, trying to do security during processing. Postulate: some node in cluster is compromised, how create record of I/O operations that have occurred. Fits with Bagit.

- Response: Bagit specification (used in LOC): Archival integrity and security integrity motivations have similar needs, although motivations are slightly different.
    - If monitoring data integrity, need repository of cryptographically secure Checksums. Use BD Bag as medium because not have to store all trusted data, can store a manifest, in lightweight manner, that you can distribute or store. If distribute files across many systems or process, use BD Bag to get on my less trusted compute system; thus, integrity of source data can be validated. Even if results tampered with. Boundary object between systems that can include data integrity checks. And systems, such as yours, can perform validating and monitoring of processes on system

Object metadata that S3 would support; Arbitrary metadata tags that can stamp files within AWS S3?

- Response: Globus not support putting MD5 directly on file in this manner. Globus just taking from point A to point B. Re: Metadata – BD Bag exists way it does because ways that are de facto standards or best practices for putting additional metadata into BD Bags and using it; checksums already in there. Globus only tracks files persistently during transfer, so no way to tag. Thus, BD Bag is complementary.

Zachary Ives, Professor and Department Chair, Computer and Information Science Department, University of Pennsylvania, *Data Provenance for Reproducibility, Integrity, and Relevance; Nan Zheng, Yi Zhang, Zhepeng Yan, Grigoris Karvounarakis, Todd J. Green, Val Tannen, Susan B. Davidson*

Professional Background: Data integration and large scale data sharing. Has worked on General purpose management tools and apply to real problems. Will address how data provenance and data integrity fit together. High level survey.

Data integrity from the broader perspective (Slide 1)

Will address connection between broader definitions of data integrity with data provenance

- Data integrity from life science perspective: Does it make sense to analyze this collection of data together?
- Data provenance: When wish to integrate and do analysis of data from multiple places, studies and sources -> Does everything come from a compatible context?

Use Cases (Slide 2)

Provenance important; Underlying everything: goals of reproducibility and unforgeability, resistance

Large scale EEG epilepsy platform.

- Goal: bring together EEG data from patients (humans and animals with epilepsy.
- First people view and annotate data; later algorithms. Then train algorithms with human annotations to perform on other data sets.
    - Issue: thousands of annotations by different tools, people, versions. Provenance of annotations: what is consensus, what is different, etc. – eventually rank quality of human annotators and ML-type annotators. Where data originates almost as important as data itself

High-throughput gene sequencing: run through analysis pipeline over span of a year to examine gene and protein expression.

- Analysis results: Produced at different times with different tools. What's compatible? What needs to be re-run?

General data science in Jupyter Notebook/JupyterLab-style environments

- How do we help scientists search "data lakes" for relevant data sources? Provenance central

Outline

What do we mean by provenance, and how is it captured?

Applications of provenance:

- Reproducibility and debugging
- Search
- Provenance and integrity

The future

Provenance for Raw Data - metadata about context (Slide 4)

Metadata that helps frame interpretation of data (what when where, why how - Who collected data? For what reason? What population?

- Important to data interpretation as well as questions of algorithmic fairness.
    - Need to know what population group and the characteristics involved.
- Might also include other contextual information (which sensor, mode, thresholds, calibration)
- Metadata is attached to data sets, manually or automatically

Issues in Contextual Metadata (Slide 5)

- How much contextual information is *available* and useful (e.g., experimental setup)?

- How much we can automate the capture (cf. EXIF data, GPS data, tablets, NFC or bar codes)?
- What standards are available? Defining metadata standards -individual communities to come up with the set of data standards.
- Tools can help, but mostly (meta) data modeling problem. Hard part is figuring out what data and metadata to capture

Focus: Derived data

Provenance for Data Products (Slide 6)

Provenance of a data product is observable if we instrument our computational environment

- Derived data – take raw data and run through workflow system or data pipeline to produce product. Want to automatically document what is going on

Conceptually

For every derived result, we want to record log of messages per activity - "trace" of: steps involved in creating the result; the inputs; the parameters.

From a tool's perspective, wish to:

- collect everything into a "provenance database management system"
- link provenance to our regulate database/file system
- be able to reason about data and provenance
- be able to verify it hasn't been tampered with

Today, many middleware systems for distributive logging

Challenge if use regular log file – difficult to do automated processing of it.

Low-level Data Provenance

- Instead build instrumentation capturing the same log information but stores in different format
- System logs all event messages, etc. happening within system (E.g. - Instrument operating system to capture all I/O, etc. to create a log, useful)
- But low-level events are typically overwhelming in scale and complexity. A log sequentializes everything, even when order doesn't matter.

Dataflow Dependencies

- Desire for something more higher-level led to dataflow driven provenance. "Lift" abstraction to show flow of data, as opposed to sequences

- Trying to capture code modules and files and their relationship (Many WF tools and systems that run script can automatically generate). Easy to use for reproducibility
    - We can annotate a file with its provenance graph, esp. from scientific workflow systems (Taverna, Kepler, Galaxy)
- High-level abstractions make it much easier to understand derivations
    - Standard formats, such as PROV, give a format for capturing entities, agents, activities
    - This provenance while useful, is too coarse grain - explains overall outputs, not specific results

Fine-Grained Provenance: Records + Operations -> Records

Equivalent SQL query will result in equivalent provenance (different execution ok)

Applications of Provenance

Fine grain provenance: useful for Reproducibility and Debugging

Reproducibility*:*

Simple provenance documents results: Can rerun computations over input to reproduce outputs (for both coarse and fine grain provenance)

More interesting question: For other data collected in similar context, can reapply same computations?

- Fine grain provenance – enables ensuring all our data produce in compatible ways, even if data has been integrated into a single repository
    - E.g. NIH NDAR data looks compatible but encoded using different subjective scales with different patient group; thus, not meaningful to compare different rate criteria from different groups.
    - When have a process, assumes input has particular structure and context.
    - Increasingly important in big data, data reuse world: Can explain answer, why 2 results differ and show why computations have similar context and compatible

Provenance and Search

How do we find data that is useful to us?

- First searched Google.
- Now "Data lake" on cloud to collect data and different versions of data. Need more sophisticated search examining version, age, other use, source - i.e., source, context and provenance

Projects:

1) Looking at dry results from different computations, use ML and user feedback to figure out which sources, tools and data are most relevant to an individual; i.e., try to learn how to predict relevance from provenance

2) Trying to build data science environments linked into the cloud that are running middleware, apache spark and big data computations – where store output in cloud and can we find other relevant data to use for training, etc.; i.e., are there other tables with similar provenance that are likely to be compatible in meaningful ways?

Tamper-Resistant Provenance

How to persistently archive data and its provenance

Cryptographic hashing

- Merkle Trees technique bind input at every level and hashes, recursively to top of tree. Certify everything that goes into tree.

Summary and Future Directions

Overview and highlight where fine grained tuple level provenance brings benefits

Thinking of data Integrity more broadly:

- Integrity of a result as it was computed from other things re-used from public data sets is increasingly important. How define when data is compatible or not for an experiment.

Integrating data:

- Problem because blurs difference that may matter in terms of interpretability of result

Provenance is glue between input data and integrated data.

Discussion

Broadly speaking, a set of tools is emerging for managing different ML configurations. Some open source companies are trying to associate ML configurations. Also, workflow provenance work adopted widely. Interest in looking at differences where they matter between individual rows. From a table search perspective, interesting connection to follow up on: classifiers trained on different version of dataset where adds/removes different features. Much room for continued fleshing out differences.

Speaker Suggestions

Data confidentiality update (May)

- Wu Chen (NIST)
- Provenance project: also, security and monitoring of operations of given data set (Alan Sill)
- Solutions: Potential speakers

    - Dan Stanzione (TACC) – treating user data as CUI (Control of Unclassified information) requires set of procedures and processes, so expenses involved. Also how does TACC view end to end picture
- Use cases: companies wish to use tools but not share data)
- Academia/Industry partnerships: Privacy, confidentiality, proprietary issues– one of biggest issues when attempting to share computing resources (e.g., shared nodes)
- AAU and APLU Summit for Accelerated Public Access to research data - Overview of data discussions from multiple points of view (e.g., who needs to part of this on campus) Sharon Broude Geva
- Implementation model/processes/framework, conditions of use (citation requirements) to protecting sensitive data – no implementation model or platform. Share current state of attempts to implement and come up with framework. (Birali Runesha, UChicago):
    - Context: Started research computing center at university. Not deal with much sensitive data. Institutional review board: some have data user agreements; also, data procured pursuant to contract; and data generated by faculty or researcher who doesn't know how to protect data.
    - University context – wish for research to be performed on AI
    - Framing question (ongoing research for implementation)
    - Privacy
        - Legal counsel – trying to put into place policies and processes; lack technical background. Developing mandatory training for privacy for researchers conducing data classification at threshold level. Regulatory compliance
        - Chief security officer - governance (non-technical)
        - Database community (privacy, computations, searches) – Vipin Chaudhary (NSF)

Data publication – Data confidentiality within systems (Alan Sill)

- Related to data confidentiality: Definition/citation requirements, conditions of use issues arise when creating data publication systems

Implications of new AI on science work (April)

- DOE AI for science series (need summary): DOE AI Town Hall series (Final report published)
- Ian Foster; Geoffrey Fox (Indiana U): streaming data and ML workshop summary

ROI and cost efficiency for academic and lab-based computing (June)

CASC Survey of over 50 research institutions (post-April presentation)

PEARC Paper -Craig Stewart, et al (Check with Alan Sill and Sharon Geva in April)

ROI, even ROI assessment, not treated consistently or thoroughly

- need best practices session and how to have them adopted
- definition and how we would want to look at it

Prerequisite for assessing delivery mechanisms

How to improve understanding of ROI throughout research enterprise

Direct integration of energy sources and computing facilities (A Sill)

Potential speaker: Andrew Grimshaw (UVA)

Roundtable

CASC: Alan Sill and Sharon Broude Geva

CASC members and government entities welcome to participate in upcoming CASC meeting on April 1-3.

DOE/SC: Richard Carlson

DOE/SC will be hosting community of Interest workshop on Future Scientific Methodologies. Co-chairs, Ian Foster and Amber Boline (Slack) for April 2020

NSF: Vipin Chaudhary

NSF call for proposals Principles and Practice of Scalable Systems. Deadline March 30th t

Meetings:

April 1 – April 3, 2020: CASC meeting, Westin Crystal City, Virginia

April 14 and 16 –DOE/SC COI Workshop: How scientific computing and data analysis will be conducted in DOE. White papers due March 5, 2020. Crystal City, Virginia)

April 29- May 1, 2020: Women in HPC Summit, Vancouver, BC

July 26 – 30, 2020, PEARC20 Meeting, Portland, OR (February 17, 2020 deadline for submissions)

Next Meeting: April 1, 2020 (12 noon ET)