Networking and Information Technology
Research and Development

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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
December 5, 2018, 12-2 pm
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

Rich Carlson	DOE/SC
Dhruva Chakravorty	TAMU
Vipin Chaudhary	NSF
Mark Day	LBL
Mariam Elsayed	DOE/SC
Ben Meekhof	UMich
Sharon Broude Geva	UMich
Dan Gunter	LBL
Florence Hudson	FDHint
Joyce Lee*	NCO
David Martin	ANL
Thomas Morton	DoD/OSD
Gilberto Pastorello	LBL
H. Birali Runesha	UChicago
Arjun Shankar	ORNL

Proceedings:
This meeting was chaired by Richard Carlson (DOE/SC). MAGIC meeting minutes from October 2018 and November 2018 were approved.

CY19 Tasking
MAGIC Report**:**
MAGIC conducted a 4 month serie on containerization systems and DevOps in CY18. In coordination with the Large Scale Networking IWG, MAGIC will generate a report on what was accomplished and learned during this series. The Co-Chairs are seeking members from the group to take the lead on these 2 tasks; specifically, to utilize meeting minutes and the presentations (available on the MAGIC website) to put together a short 4-5 page report akin to a workshop report. Anyone interested should contact the Co-chairs and Joyce Lee for additional details.

[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 Life Cycle Series Planning (February 2019 – Spring 2019)

What's changed and what's new, from generated data, re-use of data, data triage to the scientific life cycle code from large scale instruments. For example, the astronomy community is examining the data that is being used for image verification and to verify whether information is new. In the triage state, how can we capture information that is new and important and discard/archive other information? Launch room archiving was noted – how to save, also how to save in short-term archives? How to locate rapidly from short-term archive?

MAGIC members: What else should we be looking at in the data life cycle and what topics are of interest to the community?

Discussion

- Data Generation (data acquisition – getting data in a way that you can make it useful)
- Data Use/Reuse
    - Data acquisition (ensure usefulness)
    - Publishing and curating methods for library of data (how facilities make this data available)
        - Reliably assigning credit; Tracking, particularly for external curation
        - Automatically capturing metadata to facilitate sharing and information (Metadata Research Center, Drexel University – Jane Greenberg)
    - Software preservation (how data is generated, read and used)
    - Making data available/understandable
    - Data formats and standardization (e.g., tension between proprietary software/formats with scaling up)
- Triage
- Archiving (short and long term, locating data)
    - Automatic modification by routines (e.g., stream data – may suddenly find out that centers was stuck and recalibrate it, which changes data instead of noting that its "bad")
    - Calibration and stability
- Data Analysis: balance as funding agencies to understand statistical analysis. Not have to examine specific analysis routine, rather focus on the following:
    - Streaming data – distinction between analyzing streaming vs. archival data
    - ML, AI techniques vs. statistical learning techniques
    - Statistical analysis technique
    - Portfolio balancing
    - Trends
    - Data modality
    - Cross-facility view of data handling in the life cycle: on site analysis of data from experimental to observational facilities and moving data across facilities to a computational facility for analysis. Increasing interest in this balance. Discussants

noted that this may be a standalone topic, separate from data analysis as it does not focus on analysis techniques.
    - Federated data management across multiple institutions and sizes
    - Nature of analysis at different sites from inception to scalable processing)
        - How data is held in its life cycle from inception to scalable processing
    - Scalable processing
    - Sanitization (different meanings –e.g., reliability)
- Data reproducibility (e.g., analysis of streaming data)
- Data storage
    - Approaches (original generators, researchers)
- Data provenance (origination, whether data was changed)
    - Data veracity
    - Data origination: Original source, derived data; devices
    - Data versioning (distinguish, change, track changes, re-do computations, etc.)
    - Digital distributed ledger technology (DLT) / blockchain
- Data privacy/security
    - Integrity (Trusted CI)
    - Anonymization to enable sharing without releasing PII
- Data Resiliency – incorporate at every level
- Data preservation cost - tiering for different modes of storage
    - Business and cost models
- Multi-Party computation (MPC) (data protection)
- Interaction between networking and data management communities
    - Network implications of exascale, ML, higher data analysis, etc.
- Data Quality Control/Assessment
    - Scalability challenge
- Tools (e.g., Jupyter Lab) Discussed whether it should be place in the data analysis section
    - Data availability (storage and network)
    - Identity and access management (link tools)
- Data velocity (IoT, increasing integrity and security risks with use of personal devices controlling scientific instruments)
- Data Management and Implementation (e.g., interoperability, scalability)
    - Note past NSF-funded, Coalition for Academic Scientific Computation (CASC) workshop on research data management, implementation and re-use workshop. Much interest, so thinking of continuing the discussion.
- Use Cases (e.g., precision medicine –bringing multi-sourced and multi-formatted data requiring different levels of permission, smart cities)
    - Data integration, analysis and significance of multi-modality data (link to mission needs, science communities etc.) Large scale generation or distribution as part of the data life cycle.

- Interdisciplinary science

Speakers & Topics (Also discuss over email)
February (General set of talks):

- Automatically generated metadata - Jane Greenberg, Drexel University (invited)
- Data stability/variability, changing data, and impact - Deduce Project Team Speaker, LBL (Invited)
- Future lab computing in the federated computing environment /creation of federated eco-systems to perform analysis - Arjun Shankar, ORNL
- Evolution of scientific instruments /generating more data (as a machine, not as science) - Dula Parkinson, LBL (invited)

March: Use cases (specific scientific communities)

- Smart Cities - Deployment of multiple sensors and analysis - Charlie Catlett, ANL (invited)
- Precision medicine /medical imaging- Warren Kibbe (Statistics/Bioinformatics Department, Duke University)
- LHC (TBD)
- Astronomy (simulation or multi-channel observational community) – Peter Nugent, LBNL (confirmed)

April:

- Reorganize and solicit speakers (Florence Hudson will outreach in January)
- Data generation, use, re-use, archival

Use, re-use, archival to see if major topics pop up – identify speakers, can go beyond 1 month per subtopic. Virtual workshop over multiple months.

Roundtable
None

Next meeting:
February 6